Essays on asset pricing

Koijen, R.S.J.

Publication date:
2008

Essays on Asset Pricing

Essays on Asset Pricing

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof. dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op woensdag 16 april 2008 om 16.15 uur door

Ralph Sebastiaan Johannes Koijen

geboren op 23 mei 1981 te Breda.

Promotores:
prof. dr. Theo E. Nijman
prof. dr. Bas J.M. Werker

Summary

This dissertation is comprised of six papers I have written during my Ph.D. thesis. Chapter 1 is titled "When Can Life-cycle Investors Benefit from Time-varying Bond Risk Premia?" and Chapter 2 "Optimal Annuity Risk Management." Both chapters have been co-authored by Theo Nijman and Bas Werker. Chapter 3 is titled "Optimal Decentralized Investment Management," and is joint work with Jules van Binsbergen and Michael Brandt. It is forthcoming in the Journal of Finance. Chapter 4, "Mortgage Timing," is co-authored by Otto Van Hemert and Stijn Van Nieuwerburgh. It has been awarded the Glucksman First Place Research Prize for "Best Working Paper in Finance" 2007/8, Stern, NYU. Chapter 5 is titled "Predictive Regressions: A Present-value Approach," which is joint work with Jules van Binsbergen. Chapter 6 is my job market paper titled "The Cross-section of Managerial Ability and Risk Preferences."

Acknowledgements

I would like to take the opportunity to express my gratitude to several people who contributed in different ways to my thesis. First, I would like to thank my advisors Theo Nijman and Bas Werker. I benefited tremendously from their guidance and supervision throughout my Ph.D. They suggested to visit Duke University, initially for a period of three months, and fully supported me when I indicated that I preferred to extend my stay in the US; first at Duke and subsequently at NYU Stern. Our collaboration resulted into the first two chapters of this thesis, and their comments on my others papers have been very helpful and constructive.

Second, I would like to thank Michael Brandt, Ron Kaniel, and Stijn Van Nieuwerburgh for their feedback on my papers, and my job market paper in particular. Michael Brandt hosted both of my visits at Duke University, which has been an exciting period. Our joint work resulted in the the third chapter of this thesis. I met Stijn Van Nieuwerburgh during my third year as a Ph.D. student. Stijn's support and motivation during the job market process has been invaluable. I furthermore thank Geert Bekaert, Lans Bovenberg, and Frank de Jong for being part of my Ph.D. committee and for their feedback on my work.

Third, I am grateful to my other co-authors Lans Bovenberg, Juan-Carlos Rodriguez, Viorel Roscovan, Juan Rubio-Ramirez, Alessandro Sbuelz, Coen Teulings, Otto Van Hemert, and Jesus Fernandez-Villaverde for many interesting discussions that predeccesed our joint papers. In particular, I would like to thank Jules van Binsbergen with who I worked on several projects. Not only turned our collaboration out to be very productive, he also made my visits to Durham very enjoyable.

Fourth, I would like to thank the members of the department at Tilburg University,

Duke University, and NYU Stern for the many discussions, feedback on my work, and their hospitality during my visiting periods. I am also grateful to ABP Investment. I spent two days a week at ABP Investments during the first two years of my Ph.D. that have served as an important source of inspiration for several of my papers. I very much appreciate their flexibility in facilitating my visits to the US.

Tenslotte wil ik mijn ouders bedanken voor alle mogelijkheden die jullie me hebben geboden mij te ontwikkelen. Jullie motivatie, onvoorwaardelijke steun en relativeringsvermogen hebben in een bijzondere mate bijgedragen aan hetgeen ik heb bereikt. Ik bedank Mijntje voor de ruimte en motivatie die je mij hebt gegeven. Ik had dit nooit kunnen realiseren zonder jouw ongekende enthousiasme. Ik dank opa, Jeroen, Yvonne, familie en vrienden voor jullie steun en buitengewone interesse die ik heb mogen ervaren in de afgelopen jaren.

Table of Contents

Chapter 1

When Can Life-cycle Investors Benefit from Time-varying Bond Risk Premia?

Abstract

We study the consumption and portfolio choice problem for a life-cycle investor who allocates wealth to equity and bond markets. Consistent with recent empirical evidence, we accommodate time variation in bond risk premia. We analyze whether and when the investor, who has to comply with borrowing, short-sales, and liquidity constraints, can exploit variation in bond risk premia. The extent to which life-cycle constraints actually restrict dynamic bond strategies depends on the prevailing bond risk premia, the investor's age, as well as return and income realizations to date. On average, the investor is able to time bond markets only as of age 45. Tilts in the optimal asset allocation in response to changes in bond risk premia exhibit pronounced life-cycle patterns that are markedly different for the real interest rate risk premium and the inflation risk premium. We find that the economic gains realized by bond timing strategies peak around age 50 and are hump shaped over the life-cycle. The additional gains realized by implementing hedging strategies are hump shaped as well, but negligible in economic terms. To solve the model, we extend recently developed simulation-based techniques to life-cycle problems that feature multiple state variables.

1.1 Introduction

Recent studies show that long-term nominal bond returns are predictable. Cochrane and Piazzesi (2005), for instance, report that a single predictor variable, which is a linear combination of forward rates, explains up to 44% of the variation in long-term bond returns in excess of the one-year rate at an annual horizon. This suggests that investors can construct dynamic bond strategies that take advantage of this stylized fact. Indeed, Sangvinatsos and Wachter (2005) show that (unconstrained) long-term investors can realize large gains by exploiting time variation in bond risk premia. We focus on life-cycle investors instead.

These investors typically have to comply with a myriad of constraints like borrowing, short-sales, and liquidity constraints. Life-cycle constraints potentially interfere with the dynamic strategies designed to benefit from time variation in bond risk premia. We show that the way these constraints actually restrict the individual's optimal strategies depends crucially on the stage of the individual's life-cycle as well as on the realization of labor income innovations and asset returns to date. In particular, we analyze the interaction between individual constraints and time-varying, market-wide investment opportunities over the life-cycle. It is this interaction that leads to pronounced life-cycle patterns in the value derived from timing bond markets for myopic (short-term) investors, and inhibits long-term investors to act strategically by constructing hedging portfolios.

How do life-cycle constraints interfere with dynamic bond strategies? We distinguish, broadly speaking, four periods in the individual's life-cycle up to retirement. The first period (age 25 to 35) is characterized by a large stock of non-tradable human capital. The individual is not able to capitalize on future labor to increase today's consumption for reasons of adverse selection and moral hazard. The investor, therefore, consumes (almost) all income available and hardly participates in financial markets. During the second stage (age 35 to 45), the investor begins to accumulate financial wealth and allocates it almost exclusively to equity markets. Human capital resembles a (non-tradable) position in inflation-linked bonds, which reduces the individual's effective risk aversion in our model.[1] Moreover, the life-cycle constraints prohibit the individual to construct long-short portfolios that exploit attractive investment opportunities in bond markets. To invest in long-term bonds, the individual has to reduce the equity allocation. The opportunity costs of doing so are simply too high at this stage for an empirically plausible range of bond risk premia. In the third period (age 45 to 55), the individual holds substantial bond positions as the position in human capital diminished sufficiently. In addition, the individual optimally tilts the portfolio towards long-term nominal bonds in periods of high bond risk premia. These tilts are economically significant and range from -20% to $+40\%$ for a plausible range of bond risk premia. During the fourth, and final, period (age 55 to 65), the stock of human capital has largely been depleted and the individual behaves more conservatively as a result. In addition to stocks and long-term bonds, the individual holds cash positions (15% on average at age 65). The opportunity costs of reducing the cash position to tilt the portfolio to long-term bonds are smaller than the opportunity costs of cutting back on the equity allocation. This implies that

[1]There are in fact two effects at work here. Absent of idiosyncratic risk, the stock of human capital is equivalent to a (non-tradable) position in inflation-linked bonds. This endowment lowers the effective risk aversion of the investor. The idiosyncratic risk component, however, increases the effective risk aversion, see for instance Gollier and Pratt (1996) and Viceira (2001). The first effect outweighs the latter in our model, as is also found in Cocco, Gomes, and Maenhout (2005). This implies in turn that the investor acts more aggressively if the stock of human capital is large relative to financial wealth.

in periods of high bond risk premia, the investor first reduces the cash position and, only for relatively high bond risk premia, the equity allocation as well. This results in pronounced life-cycle patterns in the tilts caused by variation in bond risk premia.

To understand the interaction between constraints and time-varying investment opportunities to its fullest extent, we introduce three types of life-cycle investors. These investors are distinguished by their ability (or skill level) to successfully take advantage of time variation in bond risk premia. The first investor optimally exploits short-term variation in investment opportunities. In addition, this investor acts strategically and constructs hedging portfolios that pay off when bond risk premia turn out to be low to further smooth consumption over time. The second investor we consider takes the current bond risk premia into account in her portfolio choice, but abstracts from any strategic motives and, therefore, behaves conditionally myopically. The last, and least sophisticated, investor ignores conditioning information on bond risk premia all together. By analyzing these investors, we estimate the benefits of timing bond markets myopically and the additional benefits of acting strategically over the life-cycle.

The optimal allocation to long-term bonds and the magnitude of the tilts induced by variation in bond risk premia depend on labor income innovations and investment returns that have realized to date. After all, it is the amount of human capital relative to financial wealth that determines the individual's effective risk aversion (Bodie, Merton, and Samuelson (1992)), and thus how the individual's portfolio responds to changes in bond risk premia. A string of bad returns increases this ratio and, as a result, decreases the effective risk aversion. In similar vein, unfortunate labor income shocks decrease the ratio of human capital to financial wealth and increase the effective risk aversion. We show that past asset returns and income innovations, and in particular its interaction with the investor's age, are important determinants of the value derived from timing bond markets.

We consider the life-cycle consumption and portfolio choice problem for an individual that has access to equity and bond markets, taking into account risky and non-tradable human capital, realistic investment constraints, and time-varying bond risk premia. We calibrate our model on the basis of US data over the period January 1959 to December 2005. We provide three important contributions to the extant literature. First, we show that the individual can exploit short-term variation in bond risk premia only during later stages (as of age 45) of the life-cycle. Life-cycle constraints prevent investors to do so before that age. We decompose the total nominal bond risk premium into a real interest rate risk premium and expected inflation risk premium. Our estimates imply that the expected inflation risk premium is more persistent than the real interest rate risk premium. Consequently, tilts induced by time-varying inflation risk premia turn out to be more pronounced than those

due to changes in real interest rate risk premia. As the individual ages, two effects come into play. On one hand, the before-mentioned constraints become less restrictive due to the decreased ratio of human capital to financial wealth. On the other hand, the investor becomes more conservative due to the lower amount of human capital. The first effect dominates for the inflation risk premium and we find that tilts in the long-term bond and cash allocation are increasing over the life-cycle, but hump-shaped for equity in response to changes in the inflation risk premium. In contrast, the second effect dominates for the real rate premium, which implies that tilts in the allocation to all assets are hump shaped instead.

Second, we analyze when individuals can actually benefit from time-varying bond risk premia by introducing the three types of life-cycle investors. We find that the value of timing bond markets is around 50 basis points (bp) of certainty equivalent consumption at age 25, then increases monotonically to 90bp at age 50, and subsequently decreases to 60bp at age 60. The value of behaving strategically by constructing hedging demands is negligible (at most 2bp of certainty equivalent consumption).

Third, in deriving our results, we extend the recently developed simulation-based approach by Brandt, Goyal, Santa-Clara, and Stroud (2005). Specifically, we improve upon the optimization over the optimal asset allocation and show how to optimize over consumption in a computationally efficient way by combining the simulation-based approach with the endogenous grid method introduced by Carroll (2006). We thus show how simulation-based techniques can prove useful to solve complex life-cycle problems with multiple state variables. A separate appendix that is available online contains further details. In another application, Chapman and Xu (2007) use our approach to solve for the optimal consumption and investment problem of mutual fund managers.

Our results contrast sharply with the recent strategic asset allocation studies that argue that behaving myopically as opposed to strategically induces large utility costs on part of the investor.[2] The value of strategic behavior is lower than in related strategic asset allocation studies for three reasons. First, life-cycle constraints inhibit particular dynamic strategies that take advantage of time-varying bond risk premia. To construct hedging demands, the investor has to reduce the speculative portfolio, which is too costly for a substantial period of the individual's life-cycle. Second, human capital substitutes, in our benchmark specification for individual characteristics, for long-term bonds. As a result, long-term bonds are not part of the investment portfolio of young individuals and there is no need to hedge the corresponding investment opportunities either. Third, Wachter (2002) shows that the effective horizon in intermediate consumption problems is shorter than the last period in which the investor consumes. Wachter (2002) illustrates in a model with predictable equity

[2]See for instance Campbell, Chan, and Viceira (2003) and Sangvinatsos and Wachter (2005).

returns that the optimal strategy of a 30-year intermediate consumption investor has an effective investment horizon that is shorter than a 10-year terminal wealth investor. This implies that, at the age where our life-cycle investor holds considerable bond positions, hedging time variation in investment opportunities becomes useless.

Section 1.4 analyzes how our results modify for (i) different risk preferences of the investor, (ii) different income levels corresponding to various education levels, (iii) different correlations of income risk and asset returns, and (iv) changes in the asset menu. Although these robustness checks indicate that it is important to account for individual-specific characteristics and the asset menu available for the exact implementation of the strategies, our main results carry over to these different cases.

We focus explicitly on time-varying bond risk premia, instead of the equity risk premium, for three reasons. First, there is robust empirical evidence supporting bond return predictability, see Dai and Singleton (2002) and Cochrane and Piazzesi (2005). Predictability of equity returns is, in contrast, still heavily debated as can be deduced from recent studies by Ang and Bekaert (2007), Campbell and Yogo (2006), Cochrane (2006), Goyal and Welch (2003), Goyal and Welch (2006), Lettau and van Nieuwerburgh (2006), Pástor and Stambaugh (2006), and Binsbergen and Koijen (2007). Second, long-term bonds are of particular interest to long-term investors that are generally entitled to a stream of labor income. This (non-tradable) position in labor income is equivalent to a particular position in inflation-linked bonds together with an idiosyncratic risk component. It is therefore important to understand the demand for long-term nominal bonds in the presence of labor income, in particular when bond risk premia vary over time. Third, Lynch and Tan (2006) and Benzoni, Collin-Dufresne, and Goldstein (2006) explore the intermediate consumption problem with predictable equity returns,[3] but in absence of long-term bonds. Time-varying bond risk premia, and their interaction with individual constraints in a life-cycle framework, has not been analyzed so far and is the subject of this paper.

Our model of the financial market accommodates time-varying interest rates, inflation rates, and bond risk premia. Our model is closely related to Brennan and Xia (2002) and Campbell and Viceira (2001b), but both papers assume bond risk premia to be constant. These papers study the optimal demand for long-term bonds and show that it is optimal to hedge time variation in real interest rates, in particular for conservative investors.[4] Sangvinatsos and Wachter (2005) do allow for time variation in bond risk premia. They conclude that long-term investors that are not restricted by portfolio constraints and not endowed with non-tradable labor income can realize large economic gains by both timing bond mar-

[3]Studies that study the role of stock return predictability include Balduzzi and Lynch (1999), Brandt (1999), Campbell and Viceira (1999), Lynch and Balduzzi (2000), and Lynch (2001).

[4]See also Wachter (2003).

kets and hedging time variation in bond risk premia. This is in line with the recent asset allocation literature, which emphasizes the importance of time-varying risk premia for both tactical, short-term investors and strategic, long-term investors, see Barberis (2000), Brandt (1999), and Campbell and Viceira (1999), Campbell, Chan, and Viceira (2003), Jurek and Viceira (2007), and Wachter (2002). However, the focus of these papers is not on life-cycle investors with its inherent constraints and labor income.

Our paper also relates to the life-cycle literature, see Cocco, Gomes, and Maenhout (2005), Gomes and Michaelides (2005), Gourinchas and Parker (2002), Heaton and Lucas (1997), and Viceira (2001). These papers focus predominantly on the impact of risky, non-tradable human capital on the consumption and portfolio choice decision. These studies find (i) that there are strong age effects in the optimal asset allocation as a result of changing human capital, (ii) find binding liquidity constraints during early stages of the individual's life-cycle, (iii) find a negative relation between income risk and the optimal equity allocation, and (iv) find a high sensitivity of the optimal asset allocation to correlation between income risk and financial market risks. However, these papers restrict attention to financial markets with constant investment opportunities,[5] including constant interest and inflation rates, and bond risk premia.

Closest to our paper are presumably Munk and Sørensen (2005) and Van Hemert (2006). Both papers allow for risky, non-tradable labor income and impose standard constraints on the strategies implemented. Munk and Sørensen (2005) accommodate stochastic real rates, but assume inflation rates and bond risk premia to be constant. Van Hemert (2006) does allow for stochastic inflation rates and includes housing, but assumes risk premia to be constant. We allow for time variation in bond risk premia instead and analyze how individuals can benefit from such time variation over the life-cycle. We thus examine the interaction between exploiting time variation in investment opportunities and both realistic life-cycle constraints and changing labor income. As these constraints interfere with the optimal strategies derived for unconstrained investors without labor income, we reach different conclusions. We, therefore, integrate the long-term dynamic asset allocation and life-cycle literature.

This paper continues as follows. Section 1.2 introduces the financial market and the individual's life-cycle problem. We solve for the optimal life-cycle consumption and portfolio choice problem for the three types of investors and our benchmark specification for individual characteristics in Section 1.3. We determine in addition the economic gains of timing and/or hedging variation in bond risk premia. Section 1.4 repeats the analysis for different individual

[5]Gourinchas and Parker (2002) focus on optimal consumption policies and wealth accumulation, and abstract from optimal life-cycle portfolio choice.

characteristics and asset menus. Finally, Section 1.5 concludes. Two appendices contain further technical details. The numerical method used in this paper to solve the life-cycle problem is described in detail in the technical appendix Koijen, Nijman, and Werker (2007b).

1.2 Financial market and the individual's problem

1.2.1 Financial market

Our financial market accommodates time variation in bond risk premia. The model we propose is closely related to Brennan and Xia (2002), Campbell and Viceira (2001b), and Sangvinatsos and Wachter (2005). Brennan and Xia (2002) and Campbell and Viceira (2001b) propose two-factor models of the term structure, where the factors are identified as the real interest rate and expected inflation. Both models assume that bond risk premia are constant. Sangvinatsos and Wachter (2005) use a three-factor term structure model with latent factors and accommodate time variation in bond risk premia, in line with Duffee (2002). We consider a model with a factor structure as in Brennan and Xia (2002) and Campbell and Viceira (2001b), but generalize these models by allowing for time-varying bond risk premia.

The asset menu of the life-cycle investor includes a stock (index), long-term nominal bonds, and a nominal money market account. We start with a model for the instantaneous real interest rate, r, which is assumed to be driven by a single factor, X_1,

$$r_t = \delta_r + X_{1t},\ \delta_r > 0. \tag{1.1}$$

To accommodate the first-order autocorrelation in the real interest rate, we model X_1 to be mean-reverting around zero, i.e.,

$$\mathrm{d}X_{1t} = -\kappa_1 X_{1t}\mathrm{d}t + \sigma_1'\mathrm{d}Z_t,\ \sigma_1 \in \mathbb{R}^4,\ \kappa_1 > 0, \tag{1.2}$$

where $Z \in \mathbb{R}^{4\times 1}$ is a vector of independent Brownian motions driving the uncertainty in the financial market. Any correlation between the processes is captured by the volatility vectors.

We postulate a process for the (commodity) price index to link the real and nominal side of the economy, Π,

$$\frac{\mathrm{d}\Pi_t}{\Pi_t} = \pi_t\mathrm{d}t + \sigma_\Pi'\mathrm{d}Z_t,\ \sigma_\Pi \in \mathbb{R}^4,\ \Pi_0 = 1, \tag{1.3}$$

where π_t denotes the instantaneous expected inflation. Instantaneous expected inflation is

assumed to be affine in a second factor, X_2,

$$\pi_t = \delta_\pi + X_{2t},\ \delta_\pi > 0, \tag{1.4}$$

where the second term structure factor exhibits the mean-reverting dynamics

$$\mathrm{d}X_{2t} = -\kappa_2 X_{2t} dt + \sigma_2' \mathrm{d}Z_t,\ \sigma_2 \in \mathbb{R}^4,\ \kappa_2 > 0. \tag{1.5}$$

Concerning the stock (index), S, we postulate

$$\frac{\mathrm{d}S_t}{S_t} = (R_t + \eta_S)\,\mathrm{d}t + \sigma_S' \mathrm{d}Z_t,\ \sigma_S \in \mathbb{R}^4, \tag{1.6}$$

where R_t is the instantaneous nominal interest rate to be derived later (see (1.11)) and η_S the constant equity risk premium.

To complete our model, we specify an affine model for the term structure of interest rates by assuming that the prices of risk are affine in the real rate and expected inflation. More precisely, the nominal state price density $\phi^\$$ is given by

$$\frac{\mathrm{d}\phi_t^\$}{\phi_t^\$} = -R_t \mathrm{d}t - \Lambda_t' \mathrm{d}Z_t. \tag{1.7}$$

We assume that the time-varying prices of risk Λ_t are affine in the term structure factors X_1 and X_2,

$$\Lambda_t = \Lambda_0 + \Lambda_1 X_t, \tag{1.8}$$

and $X_t = (X_{1t}, X_{2t})$. We thus adopt the essentially affine model as proposed by Duffee (2002). In the nomenclature of Dai and Singleton (2000), the model proposed can be classified as $\mathbb{A}_0(2)$.

This specification accommodates time variation in bond risk premia as advocated by, for instance, Dai and Singleton (2002) and Cochrane and Piazzesi (2005). As we assume the equity risk premium to be constant, we have

$$\sigma_S' \Lambda_t = \eta_S, \tag{1.9}$$

which restricts Λ_1.

Given the nominal state price density in (1.7), we find for the real state price density, ϕ,

$$\begin{aligned} \frac{\mathrm{d}\phi_t}{\phi_t} &= -(R_t - \pi_t + \sigma_\Pi' \Lambda_t)\mathrm{d}t - (\Lambda_t' - \sigma_\Pi')\mathrm{d}Z_t \\ &= -r_t \mathrm{d}t - (\Lambda_t' - \sigma_\Pi')\mathrm{d}Z_t. \end{aligned} \tag{1.10}$$

As a consequence, we obtain for the instantaneous nominal interest rate

$$\begin{aligned} R_t &= r_t + \pi_t - \sigma_\Pi' \Lambda_t \\ &= \delta_R + (\iota_2' - \sigma_\Pi' \Lambda_1) X_t, \end{aligned} \tag{1.11}$$

where $\delta_R = \delta_r + \delta_\pi - \sigma_\Pi' \Lambda_0$. The conditions specified in Duffie and Kan (1996) to ensure that both nominal and real bond prices are exponentially affine in the state variables are satisfied. Hence, we find for the prices of a nominal bond at time t with a maturity $t + \tau$,

$$P(t, t+\tau) = \exp(A(\tau) + B(\tau)' X_t), \tag{1.12}$$

and for an inflation-linked bond

$$P^R(t, t+\tau) = \exp(A^R(\tau) + B^R(\tau)' X_t), \tag{1.13}$$

where $A(\tau)$, $B(\tau)$, $A^R(\tau)$, $B^R(\tau)$, and the corresponding derivations are provided in Appendix 1.A. Note that the nominal price process of a real bond is scaled by changes in the price index, i.e., the nominal price process of a real bond evolves as

$$\frac{\mathrm{d}\left(\Pi_t P^R(t, t+\tau)\right)}{\Pi_t P^R(t, t+\tau)} = \left(R_t + B^R(\tau)' \Sigma_X \Lambda_t + \sigma_\Pi' \Lambda_t\right) \mathrm{d}t + \left(B^R(\tau)' \Sigma_X + \sigma_\Pi'\right) \mathrm{d}Z_t. \tag{1.14}$$

1.2.2 Individual's preferences, labor income, and constraints

We consider a life-cycle investor who starts working at age $t_0 = 0$ and retires at age T. The individual derives utility from real consumption, C_t/Π_t, and real retirement capital, W_T/Π_T. The individual's preferences are summarized by a time-separable, constant relative risk aversion utility index. More formally, the individual solves[6]

$$\max_{(C_t, x_t) \in \mathcal{K}_t} \mathbb{E}_{t_0} \left(\sum_{t=t_0}^{T-1} \frac{\beta^t}{1-\gamma} \left(\frac{C_t}{\Pi_t} \right)^{1-\gamma} + \frac{\varphi \beta^T}{1-\gamma} \left(\frac{W_T}{\Pi_T} \right)^{1-\gamma} \right), \tag{1.15}$$

where φ governs the utility value of terminal wealth relative to intermediate consumption, β denotes the subjective discount factor, and $\mathcal{K}_t$ summarizes the constraints that have to be

[6]Recall that we normalize the price index at time 0 to one.

satisfied by the consumption and investment strategy at time t. We discuss these constraints in detail below. The fraction of wealth allocated to the risky assets at time t is indicated by x_t. The remainder, $1 - x_t'\iota$, is allocated to a nominal cash account.

The nominal, gross asset returns are denoted by R_t and the nominal, gross return on the single-period cash account is indicated by R_t^f. The dynamics of financial wealth, W_t, is then given by

$$W_{t+1} = (W_t - C_t)\left(x_t'\left(R_{t+1} - \iota R_t^f\right) + R_t^f\right) + Y_{t+1}, \tag{1.16}$$

in which Y_t denotes the income received at time t in nominal terms. The supply of labor is assumed to be exogenous.[7] For notational convenience, we formulate the problem in real terms, with small letters indicating real counterparts, i.e.,

$$c_t = \frac{C_t}{\Pi_t},\ w_t = \frac{W_t}{\Pi_t},\ r_t = \frac{R_t\Pi_t}{\Pi_{t-1}},\ r_t^f = \frac{R_{t-1}^f\Pi_{t-1}}{\Pi_t},\ y_t = \frac{Y_t}{\Pi_t}. \tag{1.17}$$

The resulting budget constraint in real terms reads

$$w_{t+1} = (w_t - c_t)\left(x_t'\left(r_{t+1} - \iota r_{t+1}^f\right) + r_{t+1}^f\right) + y_{t+1}. \tag{1.18}$$

The state variables are given by (X_t, y_t, w_t) and the control variables by (c_t, x_t), i.e., the consumption and investment choice. The set $\mathcal{K}_t = \mathcal{K}(w_t)$ summarizes the constraints on the consumption and investment policy. First, we assume that the investor is liquidity constrained, i.e.,

$$c_t \leq w_t, \tag{1.19}$$

which implies that the investor cannot borrow against future labor income to increase today's consumption. Second, we impose standard borrowing and short-sales constraints

$$x_{it} \geq 0 \text{ and } \iota' x_t \leq 1. \tag{1.20}$$

Formally, we have

$$\mathcal{K}(w_t) = \{(c, x) : c \leq w_t,\ x \geq 0, \text{ and } \iota' x \leq 1\}. \tag{1.21}$$

Note that the investor cannot default within the model as a result of these constraints.[8]

[7] Chan and Viceira (2000) relax this assumption and consider an individual who can supply labor income flexibly instead.

[8] Davis, Kubler, and Willen (2003) and Cocco, Gomes, and Maenhout (2005) accommodate costly bor-

We model real income in any specific period as

$$y_t = \exp(g_t + \nu_t + \epsilon_t), \tag{1.22}$$

with $\nu_{t+1} = \nu_t + u_{t+1}$, where $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and $u_t \sim N(0, \sigma_u^2)$. This representation follows Cocco, Gomes, and Maenhout (2005) and allows for both transitory (ϵ) and permanent (u) shocks to labor income. We calibrate g_t consistently with Cocco, Gomes, and Maenhout (2005) to capture the familiar hump-shaped pattern in labor income over the life-cycle (see Section 1.2.4 for details). In our benchmark specification, both income shocks will be uncorrelated with financial market risks. In Section 1.4, we also consider the case in which permanent income shocks, i.e., u_t, are correlated with financial market risks.

We solve for the individual's optimal consumption and investment policies by dynamic programming. The investor consumes all financial wealth in the final period, which implies that we exactly know the utility derived from terminal wealth w_T. More specifically, the time-T value function is given by

$$J_T(w_T, X_T, y_T) = \frac{\varphi w_T^{1-\gamma}}{1-\gamma}. \tag{1.23}$$

For all other time periods, we have the following Bellman equation

$$J_t(w_t, X_t, y_t) = \max_{(c_t, x_t) \in \mathcal{K}_t} \left(\frac{c_t^{1-\gamma}}{1-\gamma} + \beta \mathbb{E}_t \left(J_{t+1}(w_{t+1}, X_{t+1}, y_{t+1}) \right) \right). \tag{1.24}$$

The solution method used to solve this life-cycle problem is discussed in full detail in the technical appendix Koijen, Nijman, and Werker (2007b).

1.2.3 Types of life-cycle investors

We consider three types of investors that are distinguished by their ability to account for time variation in bond risk premia in their consumption and portfolio choice. We refer to the first investor as the 'Strategic Investor'. This investor follows the optimal life-cycle consumption and portfolio choice strategy. The Strategic Investor implements short-term (myopic) timing strategies to take advantage of the prevailing bond risk premia. In addition to myopic timing strategies, the Strategic Investor holds hedging portfolios that pay off when future bond risk premia turn out to be low to further smooth consumption over time.

The second investor we consider is the 'Conditionally Myopic Investor'. The Conditionally Myopic Investor does implement short-term bond timing strategies as well, but ignores

rowing and allow the investor to default (endogenously) within their model.

the correlation between asset returns and the state variables governing future investment opportunities. As a result, this investor will not hold hedging demands and behaves (conditionally) myopically. Formally, this implies that this investor perceives $\kappa_i = 0$, while σ_i $(i = 1, 2)$ is adapted to match the unconditional covariance matrix of the term structure factors. Furthermore, the term structure factors are uncorrelated with future asset returns.[9]

The third investor we analyze is termed the 'Unconditionally Myopic Investor'. This investor ignores time variation in bond risk premia all together. Such an investor has been studied in detail in the life-cycle literature (see for instance Cocco, Gomes, and Maenhout (2005)). Formally, the Unconditionally Myopic Investor perceives, in addition to the constraints for the Conditionally Myopic Investor, that $\Lambda_1 = 0$. The technical appendix Koijen, Nijman, and Werker (2007b) shows how the simulation-based solution method can be used to determine the optimal strategies for these three investors in a life-cycle problem.

By considering these three types of investors, distinguished by their ability to take advantage of time-varying bond risk premia, we can analyze if and, if so, when it is important to time bond markets myopically (i.e., the Conditionally Myopic Investor versus the Unconditionally Myopic Investor) and whether behaving strategically by holding hedging portfolios adds value (i.e., the Strategic Investor versus the Conditionally Myopic Investor). Our particular definition of myopia may seem unconventional since we do allow the individual's strategy to depend on current wealth relative to human capital. However, in absence of human capital, our definition coincides with the ones used in the recent strategic asset allocation literature, see Campbell and Viceira (1999), Jurek and Viceira (2007), and Sangvinatsos and Wachter (2005). We thus introduce a notion of (financial market) myopia in life-cycle models and use it in turn to analyze when life-cycle investors can exploit time variation in bond risk premia.

1.2.4 Estimation of the model

We now estimate our specification of the financial market introduced in Section 1.2.1. Section 1.2.4 describes the data that we use in estimation and we report in Section 1.2.4 the estimation results. In Section 1.2.4 we provide the individual-specific parameters of the individual's preferences and income process.

[9]There are in fact two possible approaches. On one hand, we can estimate a version of the model with restrictions on the conditioning information used, and derive the optimal policies in that case. On the other hand, we can derive the unconditional distribution from the model, which is the approach we take. Asymptotically, both approaches are equivalent. We do not expect that our choice makes much of a difference for our main conclusions.

Data

We use monthly US data as of January 1959 to December 2005 to estimate our specification of the financial market. We use six yields in estimation with 3-month, 6-month, 1-year, 2-year, 5-year, and 10-year maturities, respectively. The monthly US government yield data are the same as in Duffee (2002) and Sangvinatsos and Wachter (2005) to December 1998. These data are taken from McCulloch and Kwon up to February 1991 and extended using the data in Bliss (1997) to December 1998. We extend the time series of 1-year, 2-year, 5-year, and 10-year yields to December 2005 using data from the Federal Reserve bank of New York. The data on the 3-month and 6-month yield are extended to December 2005 using data from the Federal Reserve Bank of St. Louis.[10] Data on the price index have been obtained from the Bureau of Labor Statistics. We use the CPI-U index to represent the relevant price index for the investor. The CPI-U index represents the buying habits of the residents of urban and metropolitan areas in the US.[11] We use returns on the CRSP value-weighted NYSE/Amex/Nasdaq index data for stock returns.

Estimation

We use the Kalman filter with unobserved state variables X_{1t} and X_{2t} to estimate the model by maximum likelihood. We assume that all yields have been measured with error in line with Brennan and Xia (2002) and Campbell and Viceira (2001b). Details on the estimation procedure are in Appendix 1.B.

The relevant processes in estimation are $K_t = (X_t', \log \Pi_t, \log S_t)'$ for which the joint dynamics can be written as

$$\mathrm{d}K_t = \left(\begin{bmatrix} 0_{2\times 1} \\ \delta_\pi - \frac{1}{2}\sigma_\Pi'\sigma_\Pi \\ \delta_R + \eta_S - \frac{1}{2}\sigma_S'\sigma_S \end{bmatrix} + \begin{bmatrix} -K_X & 0_{2\times 2} \\ e_2' & 0_{1\times 2} \\ (\iota_2' - \sigma_\Pi'\Lambda_1) & 0_{1\times 2} \end{bmatrix} K_t \right) \mathrm{d}t + \Sigma_K \mathrm{d}Z_t, \tag{1.25}$$

with $\Sigma_K = (\Sigma_X', \sigma_\Pi, \sigma_S)'$, $\Sigma_X = (\sigma_1, \sigma_2)'$, and K_X a (2×2)-diagonal matrix with diagonal elements κ_1 and κ_2, respectively. An unrestricted volatility matrix, Σ_K, would be statistically unidentified and we therefore impose the volatility matrix to be lower triangular.

[10] The yield data for the period January 1999 to December 2005 are available at `http://www.federalreserve.gov/pubs/feds/2006` and `http://research.stlouisfed.org/fred2`. The data from the Federal Reserve Bank of New York are available for the cross-section of long-term yields (1-year, 2-year, 5-year, and 10-year) as of August 1971. The correlation over the period August 1971 to December 1998 of these yields with the data used in Duffee (2002) equals 99.95%, 99.97%, 99.94%, and 99.85%, respectively. The 3-month and 6-month yields are available as of January 1982. The correlation of these data over the period January 1982 to December 1998 with the data used in Duffee (2002) equals 99.96% and 99.95% for 3-month and 6-month yields.

[11] See `http://www.bls.gov` for further details.

We furthermore restrict the risk premia to obtain a single-factor term structure model for the real term structure and a two-factor model for the nominal term structure, in line with Brennan and Xia (2002) and Campbell and Viceira (2001b) in case of constant bond risk premia. To this end, we assume that the price of real interest rate risk is driven by the real interest rate only. In addition, the price of risk corresponding to the part of expected inflation risk that is orthogonal to real interest rate risk (i.e., the second Brownian motion, Z_2) is assumed to be affine in expected inflation. These restrictions imply in turn that inflation-linked bond risk premia are driven by the real rate only, while nominal bond risk premia depend on both the real rate and expected inflation.

The price of unexpected inflation risk cannot be identified on the basis of data on the nominal side of the economy alone. We impose that the part of the price of unexpected inflation risk that cannot be identified using nominal bond data equals zero. Since inflation-linked bonds have been launched in the US only as of 1997, the data available is insufficient to estimate this price of risk accurately. This restriction is in line with the recent literature, see for instance Ang and Bekaert (2007), Campbell and Viceira (2001b), and Sangvinatsos and Wachter (2005).

Formally, these constraints on the prices of risk imply

$$\Lambda_t \;=\; \Lambda_0 + \Lambda_1 X_t = \begin{bmatrix} \Lambda_{0(1)} \\ \Lambda_{0(2)} \\ 0 \\ \star \end{bmatrix} + \begin{bmatrix} \Lambda_{1(1,1)} & 0 \\ 0 & \Lambda_{1(2,2)} \\ 0 & 0 \\ \star & \star \end{bmatrix} X_t, \tag{1.26}$$

where the '$\star$' in the last row indicate that these parameters are chosen to satisfy the restriction that the equity risk premium is constant (i.e., $\sigma_S'\Lambda_0 = \eta_S$ and $\sigma_S'\Lambda_1 = 0$).

We report the estimation results in Table 1.1. The parameters are expressed in annual terms. The standard errors are computed using the outer product gradient estimator. The parameters σ_u ($u = 0.25, 0.5, 1, 2, 5, 10$) correspond to the volatility of the measurement errors of the bond yields at the six maturities used in estimation.

We briefly summarize the relevant aspects of our estimation results. We find that expected inflation is considerably more persistent than the real interest rate (i.e. $\kappa_2 < \kappa_1$), in line with Brennan and Xia (2002) and Campbell and Viceira (2001b). The (instantaneous) correlation between the real rate and expected inflation is negative (-22%). Hence, the Mundell-Tobin effect is supported by our estimates, consistent with Brennan and Xia (2002). We find that innovations in stock and bond returns are negatively correlated with inflation innovations.

We now turn to the prices of risk and implied risk premia. The equity risk premium (η_S) is estimated to be 5.4%, which reflects the historical equity risk premium. We further find that the unconditional price of real interest rate risk is slightly higher than the unconditional price of expected inflation risk, i.e., $\left|\Lambda_{0(1)}\right| > \left|\Lambda_{0(2)}\right|$. The Sharpe ratio of 5-year nominal bonds is slightly higher than for inflation-linked bonds with the same maturity (0.25 versus 0.22). For 10-year bonds, in contrast, the Sharpe ratio for inflation-linked bonds is higher (0.18 versus 0.23).

Table 1.2 reports the risk premia on both nominal and real bonds, as well as their volatilities and the inflation risk premium when the factors equal their unconditional expectation. We define the inflation risk premium as the difference in expected returns on a nominal and inflation-indexed bond with the same maturity, consistent with Campbell and Viceira (2001b). Nominal bond risk premia range from almost 60bp for a 1-year bond somewhat over 2% for a 10-year bond. Next, real bonds tend to be much safer than nominal bonds, which is caused by the fact that real bonds do not have exposure to the highly persistent expected inflation factor. The unconditional 1-year inflation risk premium equals 28 basis points and the 10-year inflation risk premium 140bp. Buraschi and Jiltsov (2005) estimate the short-term inflation risk premium to be 25bp and the long-term at 70bp in a general equilibrium setting. Campbell and Viceira (2001b) 110bp for long-term bonds.[12]

The impact of the term structure factors on bond risk premia, i.e., the time variation in prices of risk, is governed by Λ_1. Figure 1.1 presents the 5-year nominal and real bond risk premia together with the 5-year inflation risk premium for a realistic range of the real rate (0% to 4%, see the left panel) and expected inflation (0% to 8%, see the right panel). The range of the term structure variables corresponds approximately to two unconditional standard deviations around their unconditional expectation. First, we find that nominal and real bond risk premium are increasing in the real interest rate, but the inflation risk premium, which is the difference between these two, is negatively related to the real interest rate (left panel). Higher real interest rates lead to lower inflation risk premia as the exposure of real bonds to the real rate factor exceeds the exposure of nominal bonds to this factor. This latter effect is caused by the negative correlation between the real rate and expected inflation (i.e., $\sigma_{2(1)} < 0$). Second, the nominal bond risk premium increases with expected inflation, while the real bond risk premium is virtually insensitive to changes in expected inflation (right panel). Real bonds only have an exposure to the expected inflation risk premium via unexpected inflation, which is relatively small. This implies that the expected inflation risk premium is positively related to expected inflation. Buraschi and Jiltsov (2005) report the

[12] If we estimate our model up to May 2002, the 10-year inflation risk premium equals 101bp. Our estimates are thus consistent with the literature and in addition reflect the recent increase in inflation risk premia, see Koijen, van Hemert, and van Nieuwerburgh (2007) for further empirical evidence.

same relation between the real rate, expected inflation, and the inflation risk premium in a general equilibrium set-up. Third, we find that nominal bond risk premia are much more sensitive to changes in expected inflation than to changes in the real rate, which is caused by the high persistence of expected inflation discussed earlier.

Panel A of Table 1.3 presents the correlations between the assets that are possibly included in the asset menu, while Panel B of Table 1.3 reports the correlation between the risk premia on 5-year and 10-year nominal bonds and the same asset returns. These correlations are important as they drive the hedging demands held by the investor to hedge against future changes in investment opportunities. Stock returns and nominal bond returns are in turn positively correlated, consistent with Sangvinatsos and Wachter (2005). The correlations in Panel B indicate that long-term bond returns are strongly negatively correlated with bond risk premia. This implies that a long position in these bonds can be used to hedge adverse changes in bond risk premia. After all, a decrease in the risk premium on long-term nominal bonds is likely to occur jointly with a positive return on these bonds. This readily implies that the optimal allocation to long-term bonds of an unconstrained, long-term investor that is not endowed with a stream of labor income is increasing in the investment horizon, see also Sangvinatsos and Wachter (2005).

Individual-specific parameters

We now specify the parameters that govern the individual's preferences and labor income process. In our benchmark specification, we set the coefficient of relative risk aversion to $\gamma = 5$ and the subjective discount factor to $\beta = 0.96$. The investor consumes and invests from age 25 to age 65. The income process is calibrated to the model of Cocco, Gomes, and Maenhout (2005). In the benchmark specification, we focus on an individual with high school education, i.e., the "High School" individual in Cocco, Gomes, and Maenhout (2005). The variance of the transient shocks then equals $\sigma_u^2 = 0.0738$ and of the permanent shocks $\sigma_\epsilon^2 = 0.0106$. The function g_t, $t \in [25, 65]$, in (1.22) is modeled by a third order polynomial in age

$$g_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2/10 + \alpha_3 t^3/100, \tag{1.27}$$

and captures the hump-shaped pattern in labor income. The parameters are set to $\alpha_1 = 0.1682$, $\alpha_2 = -0.0323$, and $\alpha_3 = 0.0020$. The parameters α_1, α_2, and α_3 follow from Cocco, Gomes, and Maenhout (2005) and the constant is chosen so that the income level at age 25 equals \$20,000.

At retirement, we assume that all wealth in converted into an inflation-linked annuity that is priced on the basis of the unconditional expectation of the real interest rate. Koi-

jen, Nijman, and Werker (2007c) study the asset allocation problem for an individual who allocates her retirement capital to various annuity products. They show that the hedging demands before retirement induced by this retirement choice are negligible. We therefore abstract from conversion risk caused by the annuitization decision. We simplify the retirement problem further by assuming that the individual dies with probability one at age 80. This allows us to determine φ in (1.15) as the utility derived from annuitizing retirement wealth.[13,14] We have an annual decision frequency in our model.

We further analyze in Section 1.4 the impact of individual-specific characteristics, like risk preferences, education level, correlation of human capital with asset returns. We also modify the asset menu of the investor. We assume in our benchmark specification that the individual has access to the stock index, 5-year nominal bonds, and cash.[15] It is well known, see Brennan and Xia (2002) and Campbell and Viceira (2001b), that unconstrained investors will hold a long-short position in two nominal bonds to obtain the optimal exposures to both term structure factors. This strategy is infeasible in a life-cycle framework where the investor typically has to comply with borrowing and short-sales constraints. This implies that there exists at each point in time an optimal maturity of nominal bonds. Optimizing over this optimal maturity is computationally infeasible. A longer maturity of the nominal bond allows the investor to have a larger expected inflation exposure for a smaller fraction of capital allocated to long-term bonds. Very long-term bonds, however, will have again a lower real interest rate exposure due to the negative correlation between the real rate and expected inflation. In addition, bonds with maturities far beyond 10 years may not be traded liquidly. We therefore consider an alternative, realistic asset menu in which 5-year nominal bonds

[13]Specifically, denote the price of a real annuity at retirement by A_T. The time-T value function induced by annuitization is then given by

$$\sum_{t=T}^{80} \beta^{t-T} \frac{(W_T/A_T)^{1-\gamma}}{1-\gamma},$$

i.e.,

$$\varphi = \frac{A_T^{\gamma-1}}{1-\gamma} \sum_{t=T}^{80} \beta^{t-T}.$$

[14]We thus focus on the life-cycle investment and consumption problem in the pre-retirement period, consistent with for instance Benzoni, Collin-Dufresne, and Goldstein (2006). The investor's preference to save for retirement consumption is captured by the assumption that the investor derives utility from annuitized wealth up to age 80.

[15]An earlier version of this paper also contained inflation-linked bonds as part of the asset menu. However, until late in the life-cycle, investors do not include inflation-linked bonds in their optimal portfolio. The reason is that human capital can be viewed as a (non-tradable) position in inflation-linked bonds that reduces the investor's appetite to invest in inflation-linked bonds. Including these securities is, however, a straightforward extension of the model.

are replaced by 10-year nominal bonds. We finally remark that interest rate derivatives can have substantial value-added in our model. After all, interest rate derivatives may be used to create arbitrary exposures to the term structure factors without violating borrowing or short-sales constraints. However, we argue that interest rate derivatives are not available for most individual investors, on which we focus in this paper, and we therefore rule out such strategies.

1.2.5 Solution technique

Life-cycle problems generally do not allow for analytical solutions, and we use numerical techniques instead. Numerical dynamic programming is the leading solution technique in life-cycle models, see for instance Cocco, Gomes, and Maenhout (2005). This approach becomes infeasible given our number of state variables and we therefore adopt the simulation-based approach developed recently by Brandt, Goyal, Santa-Clara, and Stroud (2005). Simulation-based techniques are well suited to deal with multiple exogenous state variables that can be simulated. However, life-cycle problems are usually characterized by (at least) one endogenous state variable, which is in our case financial wealth normalized by current income, which depends on previous choices. We can therefore not simulate this variable and we specify a grid using the endogenous grid approach developed by Carroll (2006).[16] This method allows us to determine the optimal consumption policy analytically, implying that we do not have to solve numerically for the root of the Euler equation. In addition to combining these techniques, we improve upon the portfolio optimization under investment constraints and modify the simulation-based approach by Brandt, Goyal, Santa-Clara, and Stroud (2005) to account for intermediate consumption. Brandt, Goyal, Santa-Clara, and Stroud (2005) also addresses the intermediate consumption problem, but the resulting optimal consumption strategy is not ensured to be strictly positive. We modify their approach to guarantee that intermediate consumption does have this property without adding any computational complexity. An in-depth discussion of the numerical method is provided in the technical appendix Koijen, Nijman, and Werker (2007b). We there also show how simulation-based methods can be used to determine the optimal consumption and portfolio choice for the three types of investors introduced in Section 1.2.3.

[16] Barillas and Fernández-Villaverde (2006) extend the endogenous grid approach of Carroll (2006) to solve problems with multiple endogenous state variables and illustrate the efficiency gains realized by this method.

1.3 Life-cycle investors and bond risk premia

We present the optimal policies for the benchmark specification concerning preference parameters, income process, and asset menu discussed in Section 1.2.4. Section 1.3.1 analyzes the optimal portfolio choice over the life-cycle of the Strategic Investor. In Section 1.3.2, we study the optimal strategies of the Conditionally Myopic Investor and compare these strategies to those of the Strategic Investor. In Section 1.3.3, we compute the utility costs induced by the sub-optimal strategies followed by the Conditionally Myopic Investor and the Unconditionally Myopic Investor, respectively.

1.3.1 Optimal life-cycle portfolio choice for Strategic Investor

Figure 2.1 presents the optimal average allocation of financial wealth to stocks, 5-year nominal bonds, and cash over the life-cycle. The vertical axis displays the average portfolio choice, i.e., averaged over all state variables, alongside the individual's age on the horizontal axis.

Between age 25 and 35, the individual optimally allocates all financial wealth, which is little to begin with,[17] to equity. in our benchmark specification, labor income is essentially equivalent to a position in inflation-linked bonds with an idiosyncratic risk component. This non-tradable position in inflation-linked bonds imposes a large real interest rate exposure on the individual. This induces the investor to create exposure to risk factors other than the real interest rate via the investment portfolio. Since the investor is interested in the right exposures to the risk factors of total wealth (the sum of financial wealth and human capital) instead of financial wealth only, the individual's portfolio is tilted towards equity (also taking into account that the price of equity risk exceeds that of inflation risk). The individual starts to accumulate financial wealth between age 35 and 45. Nevertheless, the stock of human capital is sufficiently large for the individual to hold predominantly equity. The investor holds significant positions in long-term nominal bonds (i.e., larger than 10% on average) only as of age 45. Human capital has depleted sufficiently to reduce its effect on the portfolio choice. Between age 50 and 55, the investor has a positive demand for stocks, long-term nominal bonds, and cash as well. The reduction in human capital is equivalent to an increase in the individual's effective risk aversion coefficient (see Bodie, Merton, and Samuelson (1992)). Campbell and Viceira (2001b) and Brennan and Xia (2002) show in addition that more conservative investors prefer to hedge real interest rate risk and avoid

[17] Cocco, Gomes, and Maenhout (2005) find also that the liquidity constraint binds during the first decade of the individual's life-cycle. An unconstrained life-cycle investor optimally capitalizes future labor income to increase today's consumption as a result of the hump-shaped pattern in labor income. This is, however, prohibited by the liquidity constraint and the investor consumes (almost) all income available.

inflation risk. Absent of inflation-linked bonds in the asset menu, this implies that the investor allocates her capital to cash instead, see also Campbell and Viceira (2001b). Prior to retirement, the investor allocates on average 40% to stocks, 45% to 5-year nominal bonds, and the remaining 15% to cash.

We now illustrate how the optimal conditional allocation to the three assets responds to changes in bond risk premia. To this end, we present tilts in the optimal portfolio caused by variation in bond risk premia over the life-cycle. Figure 1.3 displays the optimal life-cycle allocation to stocks (Panel A), 5-year nominal bonds (Panel B), and cash (Panel C) for an empirically plausible range of either the real rate (left panels) or expected inflation (right panels). These figures are constructed by first regressing the optimal asset allocations along all trajectories of the simulation-based method at a certain point in time on a second-order polynomial expansion (including cross-terms) in the prevailing state variables. Note that the axes for the equity allocation are reversed 'relative to the other figures' for expository reasons.

Figure 1.3 clearly shows that tilts in the optimal allocation to any of the three assets in response to changes in bond risk premia exhibit pronounced life-cycle patterns. Up to age 35, financial wealth is allocated almost exclusively to equity, regardless of the prevailing bond risk premia. Between age 35 and 45, the individual's allocation starts to tilt to long-term nominal bonds only if expected inflation is high. Recall from Figure 1.1 that the 5-year nominal bond risk premium is increasing in both the real interest rate and expected inflation, but is much more sensitive to the latter. As the investor reduces the equity allocation as of age 45, tilts in the optimal portfolio are large and can easily range from -20% to $+40\%$ in the allocation to long-term bonds for a realistic range of inflation risk premia. Since the investor optimally holds no cash during this stage of the life-cycle, the equity allocation experiences exactly the opposite tilts once compared to the long-term bond allocation. The borrowing constraint prohibits the investor to borrow cash to take further advantage of high bond risk premia. The investor either has to reduce the equity allocation or forfeit high bond risk premia. We find that the optimal stock-bond mix is very sensitive to changes in inflation risk premia during this period. Qualitatively, the results are the same for the real rate premium in this period, but the quantitative impact is much smaller. This was suggested already by Figure 1.1. As of age 55, the optimal investment portfolio also contains cash positions due to the reduced stock of human capital. This impacts both the investor's willingness and ability to time bond risk premia. First, the investor acts more conservatively and timing risk premia adds less value as a result. Second, portfolio constraints are no longer binding, which may actually induce a larger value of timing risk premia. Interestingly, we find that the first effect dominates for the real rate premium and the second effect for the inflation risk premium. For real rate risk premia, tilts in the optimal bond allocation are hump shaped.

The change in bond risk premia is not sufficient to offset the increased effective risk aversion of the investor. For inflation risk premia, in contrast, tilts in the bond allocation steadily increase as the investor ages. In addition, to tilt the optimal portfolio towards long-term bonds, the investor first reduces the cash allocation and only for high bond risk premia, the equity allocation as well. This implies that the opportunity costs of reducing the equity allocation exceed the costs induced by reducing the cash allocation. Taken together, we find that tilts in the equity allocation in response to changes in inflation risk premia are hump shaped, but tilts in the cash allocation increase as the investor ages, consistent with the tilts in the allocation to long-term bonds.

Figure 1.4 summarizes Figure 1.3 in a compact way. We measure tilts in the optimal portfolio as the difference in the optimal allocation when the real rate and expected inflation range from minus two unconditional standard deviations to plus two unconditional standard deviations around their unconditional means. Specifically, we indicate the difference in the allocation to asset i, for an investor of age t, caused by a change in the real rate by Δ^i_{rt} and for changes in expected inflation by $\Delta^i_{\pi t}$, $i =$ stocks, bonds, or cash. All other state variables equal their unconditional expectation. For instance, to summarize the impact of the real rate on the allocation to stocks, we define

$$\Delta^{\text{stocks}}_{rt} = x^{\text{stocks}}_t(r_{\max}) - x^{\text{stocks}}_t(r_{\min}), \tag{1.28}$$

with $r_{\max} - r_{\min}$ capturing the range of the real rate. We set $r_{\max}$ equal the unconditional mean of the rate rate plus two standard deviations, and likewise $r_{\min}$ is set to the unconditional mean of the real rate minus two standard deviations. Hence, a positive Δ^i_{rt} implies an increase in the age-t allocation to asset i if the real rate increases.[18]

Figure 1.4 presents the differences of all three assets for changes in either the real rate (left panel) or expected inflation (right panel) over the life-cycle. Note that the sum of the differences at any moment of the life-cycle equals zero by construction. As discussed before, we find that tilts in the optimal asset allocation induced by variation in bond risk premia exhibit pronounced life-cycle patterns. Especially during the second part of the life-cycle, these tilts are economically significant, in particular for changes in expected inflation. For the real rate premium, tilts in the allocation to all three assets are hump shaped. In case of the inflation risk premium, tilts in the allocation to long-term bonds and cash are monotonically increasing over the life-cycle, whereas the tilts to stocks are hump shaped over the life-cycle.

In addition to changes in bond risk premia, the optimal policies depend on the realizations of asset returns and labor income innovations to date. As argued before, it is the ratio of

[18]Note that all tilts are monotonically increasing in the state variables so that this measure indeed summarizes the first-order effect of changes in the state variables presented in Figure 1.3.

human capital to financial wealth that determines the investor's effective risk aversion and portfolio choice in turn. Figure 2.2 presents tilts in the optimal allocation to stocks, 5-year nominal bonds, and cash due to an increase in financial wealth from its 25%-quantile to its 75%-quantile.[19] In other words, we subtract the optimal asset allocation at the 25%-quantile from the optimal allocation at the 75%-quantile.

We find that the optimal allocation to equity is reduced for higher levels of financial wealth, while the allocation to long-term bonds and cash increases. Higher than average levels of financial wealth reduce the impact of human capital on total wealth and increase the individual's effective risk aversion. Therefore, the individual selects a more conservative strategy. An alternative interpretation is that a string of good returns essentially shifts the individual forward in the life-cycle. Figure 2.1 shows that the optimal allocation to equity is decreasing in age, whereas the optimal allocations to long-term bonds and cash increase over the life-cycle. In contrast, lower than average levels of financial wealth level lead to a more prevalent role of human capital in the composition of total wealth. This resembles an individual who acts as in an earlier stage of her life-cycle. High asset returns or, equivalently, low income innovations tilt the optimal allocation away from stocks to long-term bonds and cash at any moment in the life-cycle. Negative asset returns, or positive labor income innovations, result in exactly the opposite effect. The effects are quantitatively most pronounced around age 50 and the impact of the level of financial wealth is in the order of magnitude of the impact of the real interest rate (see Figure 1.4).

1.3.2 Optimal life-cycle portfolio choice for Conditionally Myopic Investor

The strategies discussed so far are optimal for the Strategic Investor introduced in Section 1.2.3. This strategy is characterized by short-term bond timing strategies, especially in later stages of the life-cycle (Figure 1.3 and 1.4). In addition, the Strategic Investor holds hedging demands that pay off when bond risk premia turn out to be low. We now compare the strategies of the Strategic Investor to the Conditionally Myopic Investor. The strategy of the Conditionally Myopic Investor does take current bond risk premia into account, but ignores any dependence between future asset returns and bond risk premia. As a result, this investor holds no hedging demands.

Figure 2.3 displays the hedging demands caused by variation in bond risk premia over the life-cycle. Hedging demands are defined as the difference in the average, over the real rate and

[19] In contrast to the real interest rate and expected inflation, financial wealth is not a stationary variable. To obtain comparable tilts over the life-cycle, we compare the difference in asset allocation at different quantiles of the financial wealth distribution at a particular age.

expected inflation, optimal portfolio holdings of the Strategic Investor and the Conditionally Myopic Investor. We cannot average over financial wealth, since both strategies will result in different levels of average wealth over the life-cycle. Figure 2.2 shows that this affects the investment strategy. We therefore condition on the average financial wealth level realized by the Strategic Investor to avoid any wealth effects. This definition is, in a life-cycle framework as we have, closest to the dynamic asset allocation literature, see Campbell, Chan, and Viceira (2003) and Sangvinatsos and Wachter (2005).

The optimal hedging demands turn out to be long in 5-year nominal bonds and are financed by reducing the allocation to stocks and, in particular, cash. Long-term bond returns are negatively correlated with future bond risk premia (Table 1.3, Panel B) so that a long position in bonds position pays off exactly in those states of the economy where bond risk premia are low. However, the hedging demands are strikingly small. The axes of Figure 2.3 range only from -2% to 2%, implying that the maximum hedging demand is around 2%. This result can be explained by considering how the investor actually finances the hedging demand. The investor has to reduce the equity allocation up to, say, age 50-55 to hold hedging demands that are long in 5-year nominal bonds as a result of borrowing constraints. The opportunity costs induced by cutting back on the equity allocation turn out to be too high. The individual therefore forfeits to hedge future investment opportunities. As soon as the optimal investment strategy also includes cash positions, the investor uses this cash position to construct hedging demands. However, the individual is then already at age 55 and hedging turned useless for the remaining period. After all, Wachter (2002) shows in a model with a time-varying equity risk premium that the effective duration is substantially shortened for intermediate consumption problems as opposed to terminal wealth problems. We thus reach different conclusions about the importance of strategic behavior in our life-cycle model than related strategic asset allocation studies that consider an unconstrained investor that is not endowed with a stream of labor income.[20] Life-cycle constraints and non-tradable human capital reduce the individual's ability and willingness to time bond risk premia, which dramatically reduce the optimal hedging demands.

To summarize, we find that tilts in the optimal portfolio in response to variation in bond risk premia exhibit pronounced life-cycle patterns that are markedly different for the real rate risk premium and the inflation risk premium. We further show that hedging demands induced by time variation in bond risk premia are long in bonds, and short in equity and cash, but quantitatively very small. Since small hedging demands can still lead to large utility costs if ignored, we compute in the next section the utility costs induced by the optimal strategies of the Conditionally Myopic Investor.

[20]See for instance Campbell, Chan, and Viceira (2003) and Sangvinatsos and Wachter (2005).

1.3.3 Utility analysis

We now compare certainty equivalent consumption levels for the Strategic Investor, the Conditionally Myopic Investor, and the Unconditionally Myopic Investor introduced in Section 1.2.3. The first two investors have been discussed in the previous section. The Unconditionally Myopic investor ignores all information on bond risk for the consumption and investment policies. By comparing the value functions induced by each of these investors' strategies, we can determine the utility costs of not hedging (i.e., compare the Strategic Investor to the Conditionally Myopic Investor) and both not timing and hedging time-varying bond risk premia (i.e., compare the Strategic Investor to the Unconditionally Myopic Investor).

We are not only interested in the utility costs induced by sub-optimal strategies in the beginning of the life-cycle, but also *when* these costs are in fact realized. To illustrate, if timing bond markets is valuable only beyond age 55, we may conclude that timing bond markets is irrelevant when we consider the value function at age 25. After all, any utility gains realized after age 55 are heavily discounted by the subjective discount factor. We therefore compute the value functions at age 30, 40, 50, and 60. We can then fully understand which sub-optimal aspects are costly at which stage of the life-cycle. Furthermore, the value function will depend on financial wealth, which is determined in turn by past consumption and portfolio decisions. To compensate for differences in financial wealth that are the result of different decisions in the past, we evaluate all value functions at the before-mentioned ages at the average financial wealth level following from the strategy of the Strategic Investor. For each two strategies that we compare, we compute the utility costs of strategy 2 relative to strategy 1 as loss in certainty equivalent consumption, i.e.,

$$\text{Utility costs} = \left(\frac{J_t^2(w)}{J_t^1(w)} \right)^{\frac{1}{1-\gamma}} - 1. \tag{1.29}$$

The loss in certainty equivalent consumption corresponds to the fraction of all future consumption an investor is willing to give to replace the sub-optimal strategy by the optimal strategy of the Strategic Investor.

Figure 2.4 displays the utility costs expressed in basis points (bp) of both sub-optimal strategies (i.e., of the Conditionally Myopic and Unconditionally Myopic Investor) relative to the optimal strategy of the Strategic Investor. First, we find that the Conditionally Myopic Strategy results in a negligible welfare loss. This was already suggested by the strategies presented in the previous section. Second, we show that timing bond risk premia can generate significant value for the investor, even up to 90bp of certainty equivalent consumption around age 50-55. We further find a pronounced life-cycle pattern in the value of bond-market

timing. For an investor of age 25, the value amounts to approximately 50bp. This is caused by the fact that the first 10 years she cannot time bond markets as a result of the borrowing constraint. The utility gains to be realized in later stages are discounted by the subjective discount factor. This immediately explains the increase in utility costs as the investor ages. The investor comes closer to the moment where bond investments will make up a significant portion of the optimal portfolio and timing will turn to be valuable. However, as of age 55, the investor acts more conservatively since the stock of human capital depleted considerably. The optimal portfolio therefore contains a significant cash position and variation bond risk premia becomes less important. This explains in turn why the value of timing bond markets decreases slightly just prior to retirement.

In summary, we conclude that timing bond markets can add significantly economic value, while the value of hedging time variation in bond markets is negligible for life-cycle investors. Further, the value derived from timing bond markets exhibits a strong (hump-shaped) life-cycle pattern.

1.4 Individual characteristics and the asset menu

We now analyze how our results modify for different individual-specific characteristics, like (i) risk preferences, (ii) education level, (iii) correlation between human capital and financial markets risks, and (iv) different asset menus. We present in all cases the tilts induced by time-varying bond risk premia, the hedging demands, as well as the utility costs caused by not timing and/or hedging investment opportunities over the life-cycle.

1.4.1 Risk preferences

The coefficient of relative risk aversion equals $\gamma = 5$ in the benchmark specification. Table 1.4 presents the results for more aggressive, $\gamma = 3$, and more conservative, $\gamma = 7$, individuals. In Panel A, we present the tilts in the optimal allocation in response to changes in the real rate and expected inflation, like in Figure 1.4. We find that the more aggressive individual ($\gamma = 3$) allocates on average more to equity than the benchmark individual. As a result, the borrowing constraint binds for a longer period and the individual's optimal investment strategy is less sensitive to variation in bond risk premia. In contrast, more conservative investors ($\gamma = 7$) shift earlier into long-term nominal bonds and cash. We find that the tilts in the optimal portfolio are therefore larger than for the benchmark investor. This is the result of two effects. First, the more conservative investor is not constrained and does not allocate all financial wealth to equity. As such, the opportunity costs amount to reducing the cash position which are smaller than reducing the equity position. Second, very conservative

investors do not care about variation in risk premia and are only concerned with selection the portfolio to optimally smooth consumption over time. For our specification, the first effect still dominates for the conservative investor and tilts are more pronounced for the conservative individual.

Hedging motives (Panel B) are naturally more prevalent for the conservative investor. We indeed find that the hedging demands are increasing in the risk aversion level. The investor with a coefficient of relative risk aversion of $\gamma = 3$ holds negligible hedging demands, while the investor with $\gamma = 7$ holds hedging demands up to 4% in long-term nominal bonds around age 60. These hedging demands are quantitatively still small however.

We further quantify the costs of not hedging time variation in bond risk premia, i.e., the Conditionally Myopic Investor, or even ignoring any current information on bond risk premia, i.e., the Unconditionally Myopic Investor. Panel C presents the utility costs associated with these sub-optimal strategies, at different moments of the life-cycle.[21] The utility costs are expressed as the loss in certainty equivalent consumption in basis points. Consistent with the results in Panel A and B, we find that the value of both timing and hedging variation in bond risk premia are higher for the more conservative investor. In all cases, the value of timing exceeds the value of hedging substantially. More aggressive investors can, perhaps surprisingly, benefit less from time variation in bond risk premia. This is the result of the borrowing constraint that prevents the investor to borrow cash to invest in long-term bonds and exploit the corresponding variation in investment opportunities.

1.4.2 Education level

The education level of the benchmark individual we consider is "High School" according to the classification of Cocco, Gomes, and Maenhout (2005). We now consider different income processes as well that correspond to the "No High School" individual in Cocco, Gomes, and Maenhout (2005), for which we have $\alpha_1 = 0.1684$, $\alpha_2 = -0.0353$, and $\alpha_3 = 0.0023$ in (1.27), and the "College" individual that is characterized by $\alpha_1 = 0.3194$, $\alpha_2 = -0.0577$, and $\alpha_3 = 0.0033$ in (1.27). Higher education levels correspond to higher expected income growth. This implies that the borrowing constraints during early stages of the life-cycle bind for a longer period as the individual prefers to capitalize future income for the purpose of consumption smoothing. We thus expect that individuals with higher education levels are even more restricted in their ability to time bond markets, especially during early stages of

[21] As before, we correct for any differences in financial wealth that arise due to different strategies before the moment at which we compare value functions. This implies, in turn, that the utility costs can be interpreted as the loss in certainty equivalent consumption from following a particular sub-optimal strategy from that age onwards. All value functions are evaluated in the average financial wealth level that follows from the strategy implemented by the Strategic Investor.

the life-cycle. Table 1.5 presents the main results.

We find that tilts in the optimal portfolio are larger for individuals in the lowest education group, see Panel A, in particular up to age 45. The borrowing and liquidity constraints during the first two decades are less restrictive for individuals with low education levels and they can therefore benefit from variation in bond risk premia. Individuals with high education levels allocate most financial wealth to equity as their income pattern is very steep and thus more back-loaded. This effect is particularly pronounced up to age 45-50. From that age onwards, tilts in the optimal portfolios are very similar across all education groups. Panel B indicates that the hedging demands are marginally larger for individuals within the "No High School" group, but these effects are again small. This is furthermore apparent from Panel C in which we show that the loss in certainty equivalent consumption as a result of not hedging variation in investment opportunities is negligible. The loss of not timing bond risk premia does result in large utility costs on part of the individual, and these costs are higher for groups with lower education. This is in particular the case up to age 50. In summary, individuals with higher education levels on average experience higher income growth. This restricts the dynamic bond strategies even further and reduces the value-added of timing bond markets. Finally, note that at age 60, when the role of human capital in total wealth is small, the utility costs of not timing bond markets are equal across education levels at 58bp.

1.4.3 Correlation between income and financial market risks

We now analyze the impact of correlation between labor income uncertainty and asset returns. Specifically, we consider the case where permanent income innovations (u) are correlated with equity returns. This correlation likely depends on the individual's occupation, education level, age, and gender. Cocco, Gomes, and Maenhout (2005) estimate the correlation between permanent labor income shocks and stock returns between -1% and 2%. Heaton and Lucas (2000) report estimates between -7% and 14%. Munk and Sørensen (2005) provide an estimate for this correlation of 17%. Finally, Davis and Willen (2000) report estimates between -25% and 30% for the correlations between a broad equity index and labor income innovations. Regarding the correlation between labor income risk and industry-specific equity risk, the correlation ranges between -10% and 40% in their estimates, depending on the individual's education level, age, and gender. We follow Viceira (2001) instead and consider an individual whose labor permanent labor income innovations exhibit a correlation of 25% with the stock index returns.[22]

[22] Alternative labor income dynamics have been proposed in the literature. Benzoni, Collin-Dufresne, and Goldstein (2006) show that even though labor income and stock markets have a low correlation in the short run, this correlation increases at longer horizons. Benzoni, Collin-Dufresne, and Goldstein (2006) subsequently analyze a model in which labor income and stock market prices are co-integrated. Storesletten,

The positive correlation of permanent income shocks with equity returns has two effects, namely the substitution effect and the value effect. The substitution effect implies that non-tradable labor income substitutes for investments in particular assets in the investment portfolio. In the benchmark specification, human capital is essentially a position in inflation-indexed bonds, if we abstract from idiosyncratic risk. This tilts the portfolio towards equity and away from long-term nominal bonds. We now consider a case in which labor income innovations and equity returns are positively correlated. This implies that the investor will reduce the equity allocation (see also Viceira (2001)) and possibly allows for a larger impact of variation in bond risk premia. The value effect refers to the change in the value of human capital for different correlations of income innovations with financial risks that are prices. For instance, if labor income is perfectly correlation with equity returns, future income is effectively discounted at the expected real return on equity instead of the yield on an inflation-linked bond. This reduces the ratio of human capital to financial wealth and increases the effective risk aversion in turn. The results are presented on the left-hand side of Table 1.6. If we compare the tilts in the optimal allocation of Table 1.6 to the ones in Figure 1.4 corresponding to the benchmark specification of uncorrelated income innovations, we indeed find larger tilts in the beginning of the life-cycle. The positive correlation of income innovations crowds out the equity investment and allows the investor to time bond markets instead. In similar vein, the hedging demands (Panel B) increase slightly, which again reflects the more important role of bond market timing when labor income and stock returns are positively correlated. In utility terms (Panel C), we find that the utility costs for young individual of not timing bond markets increases substantially. The value of hedging remains negligible.

1.4.4 Alternative asset menus

We now consider the impact of the composition of the asset menu. The results presented so far apply to an asset menu that is comprised of stocks, 5-year nominal bonds, and cash. The right-hand side of Table 1.6 displays the results for an asset menu in which the 5-year nominal bond is replaced by 10-year nominal bonds. The important differences between 5-year and 10-year nominal bonds are that the latter have a larger exposure to expected inflation risk and a smaller exposure to the real interest rate risk. The smaller exposure to the real rate factor is a result of the negative correlation between the real rate and expected inflation ($\sigma_{2(1)} < 0$). In fact, it turns out that the nominal bond risk premium is decreasing in the real interest rate. 10-year nominal bonds thus allow the investor to create a larger

Telmer, and Yaron (2004) and Lynch and Tan (2006) show that idiosyncratic labor income risk varies considerably over the business cycle. Lynch and Tan (2006) provide furthermore evidence that income growth is higher during economically good times.

expected inflation exposure with a smaller allocation of wealth, without incurring a large real rate exposure which is undesirable given the stock of human capital.

Panel A of Table 1.6 shows that the optimal allocation to all assets is sensitive to changes in either the real rate or expected inflation. For long-term bonds, these tilts are positive for high levels of expected inflation, but negative for high levels of the real rate, consistent with the exposures that bond risk premia have to these factors. Next, Panel B indicates that the hedging demands are somewhat larger when 10-year nominal bonds are part of the asset menu. 10-year bonds can be used to create larger exposures to expected inflation risk and are thus more useful to time bond markets. This improved ability to time bond markets translates into larger hedging demands as well (see Panel B). Panel C compares the strategies of the two sub-optimal investors, i.e., the Conditionally Myopic and the Unconditionally Myopic, to the strategy of the Strategic Investor. We find that the utility costs induced by not timing bond risk premia triples if 10-year nominal bonds are included in the asset menu. In addition, the costs of acting myopically instead of strategically increases, but is still dramatically smaller than the utility costs of not timing bond risk premia. We conclude that the results are qualitatively robust to the bond maturity, but also that the costs induced by sub-optimal strategies and the actual strategies required to take advantage of time variation in bond risk premia do depend on the composition of the asset menu.

1.5 Conclusions

We solve a realistic life-cycle consumption and investment problem for an investor who has access to stocks, long-term nominal bonds, and cash. Consistent with recent empirical evidence, we accommodate time variation in bond risk premia. Life-cycle investors typically have to comply with borrowing, short-sales, and liquidity constraints. We ask the question if (and, if so, when) a life-cycle investor can actually exploit this stylized fact of bond returns. We find that investors can indeed benefit significantly from timing bond markets, but this value is predominantly realized during later stages of the life-cycle. Moreover, the utility gains are induced almost fully by myopic timing strategies, and not by hedging strategies.

We decompose the total nominal bond risk premium into the real interest rate and expected inflation premium. We show that tilts in the optimal asset allocation as a result of changes in real interest rates are smaller than for changes in expected inflation. This is the result of the higher persistence in expected inflation. Tilts in the optimal allocation exhibit pronounced life-cycle patterns. For the real rate, tilts in all assets are hump shaped. For the inflation risk premium, in contrast, tilts in the allocation to long-term nominal bonds and cash are monotonically increasing in age, but hump shaped for equity. To further analyze

the economic costs induced by sub-optimal investment strategies, we introduce three types of life-cycle investors that are distinguished by their ability to exploit variation in bond risk premia. We first compare the investor who implements the optimal strategies to an investor who does incorporate current bond risk premia into her policies, but abstracts from hedging variation in bond risk premia. We find that the difference in the average optimal portfolio holdings, the so-called hedging demands, are small. In addition, the utility costs of behaving conditionally myopic are negligible. Subsequently, we consider an investor who ignores conditioning information on bond risk premia all together. We find that the utility costs associated with this strategy are typically large, and are hump-shaped over the life-cycle.

We check robustness of our results to risk preferences, education level, different correlations of income risk and asset returns, and the exact composition of the asset menu. We confirm once more that it is indeed important to account for individual-specific characteristics and the asset menu available to implement the strategies. To derive the results, we extend recently developed simulation-based methods to solve for the optimal consumption and portfolio choice. This allows us to solve life-cycle problems with a large number of state variables and multiple assets as we have.

This paper can be extended in various directions. First, we abstract from housing in our life-cycle problem. If the investor finances the house via a mortgage, there are essentially two types, namely fixed-rate mortgages (FRM) and adjustable-rate mortgages (ARM). FRMs can be interpreted as short positions in long-term bonds and ARMs as short positions in cash. This additional flexibility may allow the investor to improve upon the exposures to the term structure factors. A second extension would be to allow for an endogenous retirement decision and analyze how this interacts with the prevailing investment opportunities at bond markets. Third, we have assumed now that the household is sufficiently knowledgable to implement dynamic bond strategies. If this is not the case, see for instance Campbell (2006) and Agarwal, Driscoll, Gabaix, and Laibson (2007), the investment decision needs to be delegated to specialized portfolio managers. This will lead to additional inefficiencies.

1.A Pricing nominal and inflation-linked bonds

We derive the nominal prices of both nominal and inflation-linked bonds in the financial market described in Section 1.2, following the results on affine term structure models in, for instance, Duffie and Kan (1996) and Sangvinatsos and Wachter (2005).

To that extent, we assume that both nominal and inflation-linked bond prices are smooth functions of time and the term structure factors X. Denote the price of a nominal bond at time t that matures at time T by $P(X_t, t, T)$. Since nominal bonds are traded assets, we must have that $\phi_t^{\$} P(X_t, t, T)$ is a martingale, where $\phi^{\$}$ is given in (1.7). This implies

$$-P_X K_X X + P_t + \frac{1}{2} tr\left(\Sigma_X' P_{XX} \Sigma_X\right) - RP - P_X' \Sigma_X \Lambda = 0, \tag{1.30}$$

where the subscripts of P denote partial derivatives with respect to the different arguments and We summarize the financial market for future reference. Denote the state vector containing both term structure factors by X_t and we use

$$X_t = \begin{bmatrix} X_{1t} \\ X_{2t} \end{bmatrix}, \ \Sigma_X = \begin{bmatrix} \sigma_1' \\ \sigma_2' \end{bmatrix}, \ K_X = \begin{bmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{bmatrix}. \tag{1.31}$$

so that we have

$$\mathrm{d}X_t = -KX_t \mathrm{d}t + \Sigma_X \mathrm{d}Z_t. \tag{1.32}$$

Subsequently, Duffie and Kan (1996) have shown that in this case, when the diffusion of the state variables under the risk neutral measure is affine in the state variables and the instantaneous nominal short rate is affine in the state variables, we obtain nominal bond prices that are exponentially affine in the state variables, i.e.,

$$P(X, t, t+\tau) = \exp\left(A(\tau) + B(\tau)' X\right). \tag{1.33}$$

Substituting this expression in (2.34) and matching the coefficients on the constant and the state variables X, we obtain the following set of ordinary differential equations

$$\begin{aligned} A'(\tau) &= -B(\tau)' \Sigma_X \Lambda_0 + \frac{1}{2} B(\tau)' \Sigma_X \Sigma_X' B(\tau) - \delta_R, & (1.34) \\ B'(\tau) &= -\left(K_X' + \Lambda_1' \Sigma_X'\right) B(\tau) - \left(\iota_2 - \Lambda_1' \sigma_\Pi\right), & (1.35) \end{aligned}$$

where ι_2 denotes a two dimensional vector of ones. We also have the boundary conditions

$$A(0) = 0, \ B(0) = 0. \tag{1.36}$$

The ODEs can be solved in closed form, see for instance Dai and Singleton (2002). This leads to

$$\begin{aligned} B(\tau) &= \left(K_X' + \Lambda_1' \Sigma_X'\right)^{-1} \left[\exp\left(-\left(K_X' + \Lambda_1' \Sigma_X'\right)\tau\right) - I_{2\times 2}\right] \left(\iota_2 - \Lambda_1' \sigma_\Pi\right), & (1.37) \\ A(\tau) &= \int_0^\tau A'(s) ds, & (1.38) \end{aligned}$$

where $I_{2\times 2}$ denotes the two by two identity matrix.

For inflation-linked bonds, the derivation is slightly more involved. In this case, the nominal price of a real bond is denoted by the product $\Pi_t P^R(X, t, T)$. The martingale property of $\phi_t^{\$} \Pi_t P^R(X_t, t, T)$ leads to

$$-P_X^R K_X X + P_t^R + \frac{1}{2} tr\left(\Sigma_X' P_{XX}^R \Sigma_X\right) - (R - \pi + \sigma_\Pi' \Lambda) P^R + P_X^{R\prime} \Sigma_X \left(\sigma_\Pi - \Lambda\right) = 0, \tag{1.39}$$

Since we postulate that the instantaneous expected inflation is affine in the state variables, the price process corresponding to holding a real bond is also affine under the risk-neutral measure and we conjecture

$$P^R(X, t, t+\tau) = \exp\left(A^R(\tau) + B^R(\tau)' X\right), \tag{1.40}$$

implying that (2.37) boils down to

$$-B^R(\tau)'K_X X - A'^R(\tau) - B'^R(\tau)'X + \frac{1}{2}B^R(\tau)'\Sigma_X\Sigma_X'B^R(\tau) - r + B^R(\tau)'\Sigma_X\left(\sigma_\Pi - \Lambda\right) = 0. \tag{1.41}$$

We again match the coefficients on the constant and the state variables X, which leads to the following set of ordinary differential equations

$$\begin{aligned} A'^R(\tau) &= \frac{1}{2}B^R(\tau)'\Sigma_X\Sigma_X'B^R(\tau) - \left(\delta_R - \delta_\pi + \sigma_\Pi'\Lambda_0\right) + \left(B^R(\tau)'\Sigma_X\right)\left(\sigma_\Pi - \Lambda_0\right); \\ B'^R(\tau) &= -\left(K_X' + \Lambda_1'\Sigma_X'\right)B^R(\tau) - e_1, \end{aligned} \tag{1.42}$$

where e_i denotes the i-th unit vector. Again we can find easily an expression for $B^R(\tau)$, i.e.,

$$B^R(\tau) = \left(K_X' + \Lambda_1'\Sigma_X'\right)^{-1}\left(\exp\left(-\left(K_X' + \Lambda_1'\Sigma_X'\right)\tau\right) - I_{2\times 2}\right)e_1. \tag{1.43}$$

1.B Estimation procedure

Our estimation procedure in closely related to Sangvinatsos and Wachter (2005). The main difference is that we allow all yields to be measured with error, following de Jong (2000), Brennan and Xia (2002), and Campbell and Viceira (2001b). However, we assume that the measurement errors are independent, both sequentionally and cross-sectionally. The continuous time equations underlying the financial market in Section 1.2 can be written as

$$\begin{aligned} \mathrm{d}\begin{bmatrix} X_t \\ \log\Pi_t \\ \log S_t \end{bmatrix} &= \left(\begin{bmatrix} 0_{2\times 1} \\ \delta_\pi - \frac{1}{2}\sigma_\Pi'\sigma_\Pi \\ \delta_R + \eta_S - \frac{1}{2}\sigma_S'\sigma_S \end{bmatrix} + \begin{bmatrix} -K_X & 0_{2\times 2} \\ e_2' & 0_{1\times 2} \\ (\iota_2' - \sigma_\Pi'\Lambda_1) & 0_{1\times 2} \end{bmatrix}\begin{bmatrix} X_t \\ \log\Pi_t \\ \log S_t \end{bmatrix}\right)\mathrm{d}t + \begin{bmatrix} \Sigma_X \\ \sigma_\Pi' \\ \sigma_S' \end{bmatrix}\mathrm{d}Z_t \\ &= \left(\Theta_0 + \Theta_1 K_t\right)\mathrm{d}t + \Sigma_K \mathrm{d}Z_t. \end{aligned} \tag{1.44}$$

with

$$K_t = \begin{bmatrix} X_t \\ \log\Pi_t \\ \log S_t \end{bmatrix}. \tag{1.45}$$

As K_t follow a standard multivariate Ornstein-Uhlenbeck process, we may write the exact discretization (see, e.g., Bergstrom (1984) and Sangvinatsos and Wachter (2005))

$$K_{t+h} = \mu^{(h)} + \Gamma^{(h)}K_t + \varepsilon_{t+h}, \tag{1.46}$$

where $\varepsilon_{t+h} \overset{i.i.d.}{\sim} N\left(0_{4\times 1}, \Sigma^{(h)}\right)$ for appropriate $\mu^{(h)}$, $\Gamma^{(h)}$, and $\Sigma^{(h)}$ which we derive below. To derive the discrete time parameters, we consider the eigenvalue decomposition[23] $\Theta_1 = UDU^{-1}$. The parameters in the VAR(1) - model relate to the structural parameters via

$$\begin{aligned} \Gamma^{(h)} &= \exp\left(\Theta_1 h\right) = U\exp\left(Dh\right)U^{-1}; \\ \mu^{(h)} &= \left[\int_t^{t+h}\exp\left(\Theta_1\left[t+h-s\right]\right)ds\right]\Theta_0 \\ &= UFU^{-1}\Theta_0, \end{aligned} \tag{1.47}$$

where F is a diagonal matrix with elements $F_{ii} = h\alpha\left(D_{ii}h\right)$, with

$$\alpha(x) = \frac{\exp(x) - 1}{x},$$

[23]Note that, since K_X is a diagonal matrix, the eigenvalues of Θ_1 are given by κ_1, κ_2, and 0 (with multiplicity two). Recall that a square matrix is diagonalizable if and only if the dimension of the eigenspace of every eigenvalue equals the multiplicity of the eigenvalue. This condition is satisfied for Θ_1.

and $\alpha(0) = 1$. The derivation of $\Sigma^{(h)}$ is a bit more involved. We have

$$\begin{aligned} \Sigma^{(h)} &= \int_t^{t+h} \exp\left(\Theta_1\left[t+h-s\right]\right)\Sigma_K\Sigma_K'\exp\left(\Theta_1\left[t+h-s\right]\right)\mathrm{d}s \\ &= UVU', \end{aligned} \tag{1.48}$$

where V is a matrix with elements

$$\begin{aligned} V_{ij} &= \left[\int_t^{t+h} \exp\left(D\left[t+h-s\right]\right)U^{-1}\Sigma_K\Sigma_K'\left(U^{-1}\right)'\exp\left(D\left[t+h-s\right]\right)ds\right]_{ij} \\ &= \left[U^{-1}\Sigma_K\Sigma_K'\left(U^{-1}\right)'\right]_{ij}\int_t^{t+h}\exp\left(\left[D_{ii}+D_{jj}\right]\left[t+h-s\right]\right)\mathrm{d}s \\ &= \left[U^{-1}\Sigma_K\Sigma_K'\left(U^{-1}\right)'\right]_{ij} h\alpha\left(\left[D_{ii}+D_{jj}\right]h\right). \end{aligned} \tag{1.49}$$

Using data on six yields, stock returns, and inflation, we estimate the model using the Kalman filter. The transition equation is given by (2.42). We assume that all yields are measured with measurement error, in line with de Jong (2000), Brennan and Xia (2002), and Campbell and Viceira (2001b). On the other hand, Duffee (2002) and Sangvinatsos and Wachter (2005) select certain maturities and fit these exactly, which is tantamount to identifying the factors. In line with these papers,[24] we assume the measurement to be Gaussian and independent of the innovations in the transition equation. The likelihood can subsequently be constructed using the error-prediction decomposition, see for instance Harvey (1989).

[24]Notably, de Jong (2000) allows for cross-sectional correlation between the measurement errors.

1.C Tables and figures

Parameter	Estimate	Standard error	Parameter	Estimate	Standard error
Expected inflation: $\pi_t = \delta_\pi + X_{2t}$					
δ_π	3.48%	1.35%			
Nominal interest rate: $R_t = \delta_R + (\iota - \sigma'_\Pi \Lambda_1)' X_t$					
δ_R	5.18%	1.40%			
Dynamics term structure factors: $\mathrm{d}X_t = -K_X X_t \mathrm{d}t + \Sigma_X \mathrm{d}Z_t$					
κ_1	1.254	0.184	σ_1	1.91%	0.09%
κ_2	0.165	0.066	$\sigma_{2(2)}$	1.35%	0.03%
			$\sigma_{2(1)}$	-0.30%	0.05%
Dynamics inflation: $\mathrm{d}\Pi_t/\Pi_t = \pi_t \mathrm{d}t + \sigma'_\Pi \mathrm{d}Z_t$					
$\sigma_{\Pi(1)}$	0.14%	0.05%	$\sigma_{\Pi(3)}$	1.11%	0.03%
$\sigma_{\Pi(2)}$	0.16%	0.05%			
Dynamics equity index: $\mathrm{d}S_t/S_t = (R_t + \eta_S)\mathrm{d}t + \sigma'_S \mathrm{d}Z_t$					
η_S	5.42%	2.51%	$\sigma_{S(3)}$	-1.68%	0.66%
$\sigma_{S(1)}$	-1.62%	0.54%	$\sigma_{S(4)}$	14.85%	0.32%
$\sigma_{S(2)}$	-2.27%	0.64%			
Prices of risk: $\Lambda_t = \Lambda_0 + \Lambda_1 X_t$					
$\Lambda_{0(1)}$	-0.250	0.083	$\Lambda_{1(1,1)}$	-36.334	11.370
$\Lambda_{0(2)}$	-0.228	0.052	$\Lambda_{1(2,2)}$	-10.073	4.905
Volatility measurement error: σ_u $(u = 0.25, 0.5, 1, 2, 5, 10)$					
$\sigma_{0.25}$	0.47%		σ_2	0.12%	
$\sigma_{0.5}$	0.22%		σ_5	0.00%	
σ_1	0.02%		σ_{10}	0.22%	

Table 1.1: Estimation results for the financial market in Section 1.2
The financial market model described in Section 1.2 is estimated by Maximum Likelihood using monthly US data on six bond yields, inflation, and stock returns over the period from January 1959 up to December 2005. The bond maturities used in estimation are 3-month, 6-month, 1-year, 2-year, 5-year, and 10-year. All yields are assumed to measured with error. The first term structure factor (X_1) corresponds to the real interest rate and the second factor (X_2) to expected inflation. The CPI-U index is used to represent the relevant price index for the investor. Stock returns are based on the CRSP value-weighted NYSE/Amex/Nasdaq index. The parameters are expressed in annual terms. The standard errors are determined using the outer product gradient estimator.

Panel A: Risk premia			
Maturities	1-year	5-year	10-year
Nominal bonds	0.57%	1.59%	2.18%
Inflation-linked bonds	0.29%	0.73%	0.78%
Inflation risk premium	0.28%	0.86%	1.40%
Panel B: Volatilities			
Maturities	1-year	5-year	10-year
Nominal bonds	1.71%	6.29%	11.81%
Inflation-linked bonds	1.73%	3.27%	3.45%

Table 1.2: Risk premia and volatilities

Panel A presents the risk premia on 1-year, 5-year, and 10-year nominal and inflation-linked bonds using the estimation results in Table 1.1. The inflation risk premium is defined as the difference in expected returns on a nominal and inflation-indexed bond with the same maturity. The term structure factors (X_t) equal their unconditional expectation. Panel B displays the volatilities of the same bonds. The risk premia and volatilities are expressed in annual terms.

Panel A: Correlation of asset returns			
	Stock return	5-year nom. bond return	10-year nom. bond return
Stock return	1		
5-year nom. bond return	0.159	1	
10-year nom. bond return	0.130	0.965	1
Panel B: Correlation of asset returns with risk premia			
	Stock return	5-year nom. bond return	10-year nom. bond return
Risk premium on 5-year nom. bond	-0.174	-0.981	-0.896
Risk premium on 10-year nom. bond	-0.026	-0.611	-0.796

Table 1.3: Correlations between asset returns and risk premia

Panel A presents the correlations between stock returns, 5-year nominal bonds, and 10-year nominal bonds using the estimates reported in Table 1.1. Panel B depicts the correlation between the same asset returns and either 5-year or 10-year nominal bond risk premia.

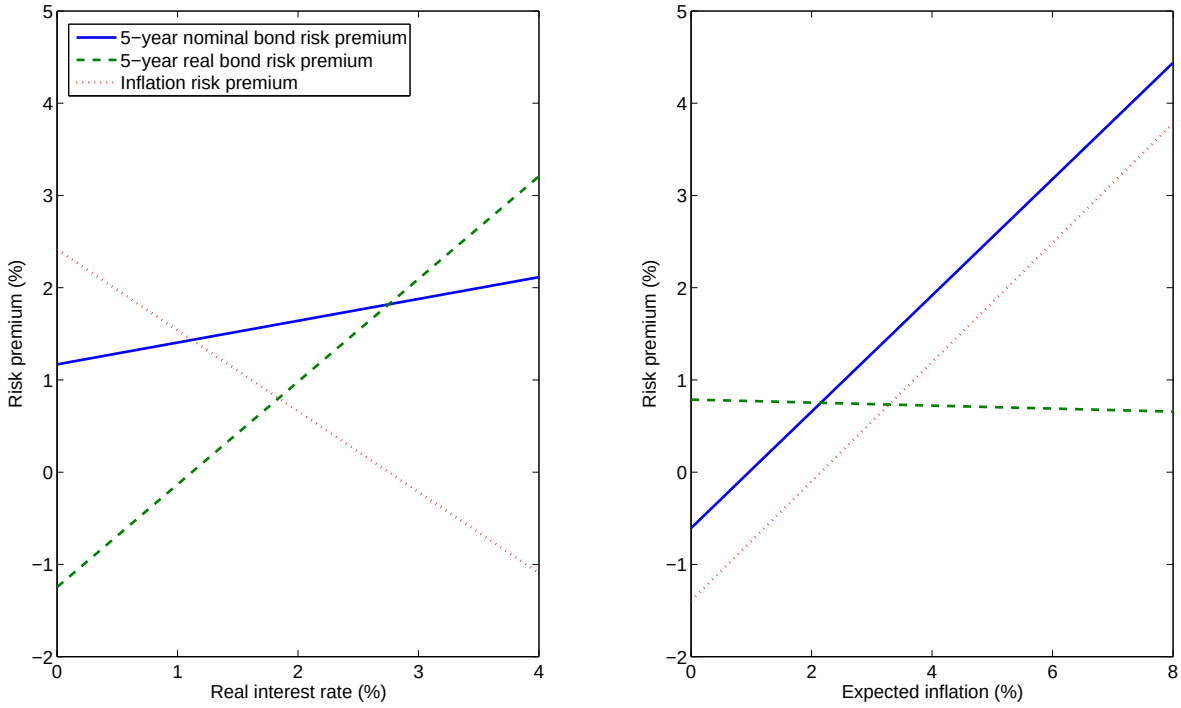


Figure 1.1: Time-variation in risk premia

The left panel presents the 5-year nominal bond risk premium, 5-year real bond risk premium, and 5-year inflation risk premium for real interest rates (horizontal axis) ranging between 0% and 4%. The right panel displays the same risk premia for expected inflation rates ranging between 0% and 8%. The range of the term structure variables corresponds approximately to two unconditional standard deviations around their unconditional expectations. The risk premia are expressed in annual terms.

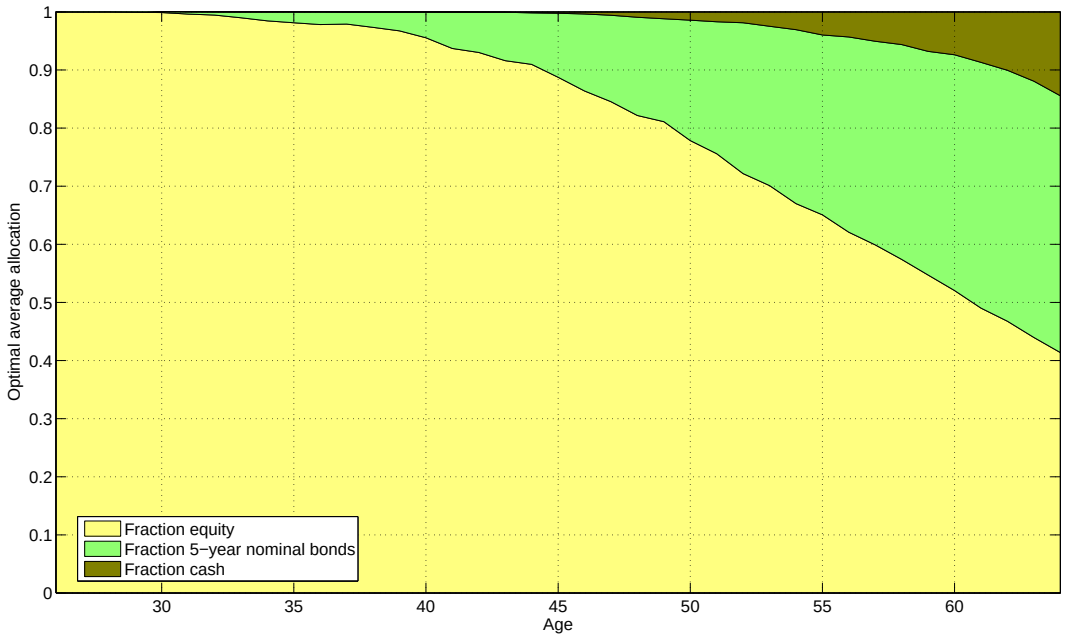


Figure 1.2: Optimal average asset allocation to stocks, 5-year nominal bonds, and cash

Optimal asset allocation over the individual's life-cycle, averaged over all state variables. The individual allocates capital to stocks, 5-year nominal bonds, and cash. The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. The vertical displays the average allocation and the horizontal axis indicates the individual's age.

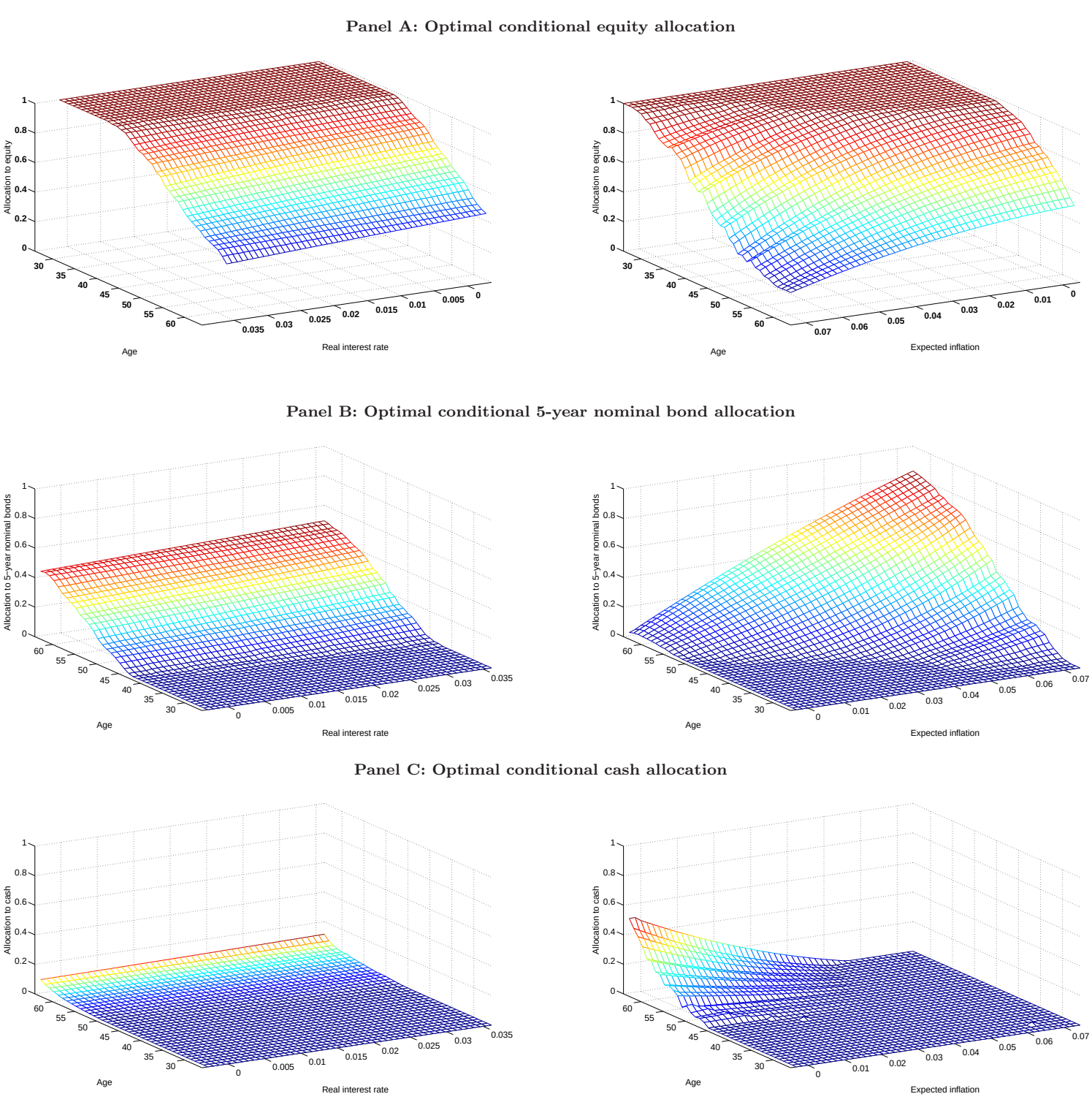


Figure 1.3: Optimal conditional asset allocation to stocks, 5-year nominal bonds, and cash
Optimal asset allocation over the individual's life-cycle conditional on either the real rate (left panels) or expected inflation (right panels). The individual allocates financial wealth to stocks (Panel A), 5-year nominal bonds (Panel B), and cash (Panel C). The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. Note that the axes for the equity allocation are reversed relative to the other figures for expository reasons. The vertical axes displays the conditional allocation and the horizontal axes the individual's age, and either the real rate or expected inflation.

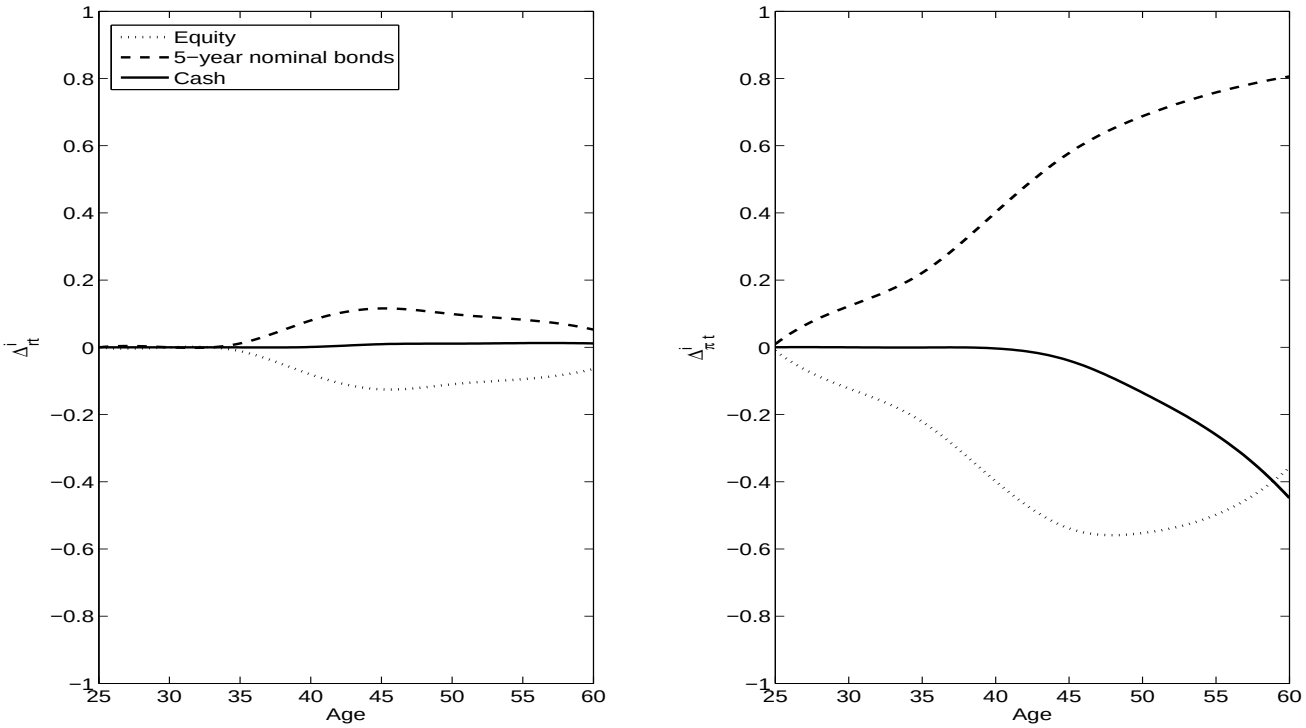


Figure 1.4: Tilts in the optimal asset allocation induced by bond risk premia
Tilts in the optimal allocation to stocks, 5-year nominal bonds, and cash in response to changes in either the real interest rate (left panel) and expected inflation (right panel). We present the difference Δ^i_{jt}, in which $j = r$ or π and i indicates the asset, in the optimal allocation between a high and a low value of the state variable. The state variables range from minus two unconditional standard deviations to plus two unconditional standard deviations around their unconditional mean. The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. The vertical axes display the difference in the optimal allocation and the horizontal axes the individual's age. We present five-year averages to summarize the results.

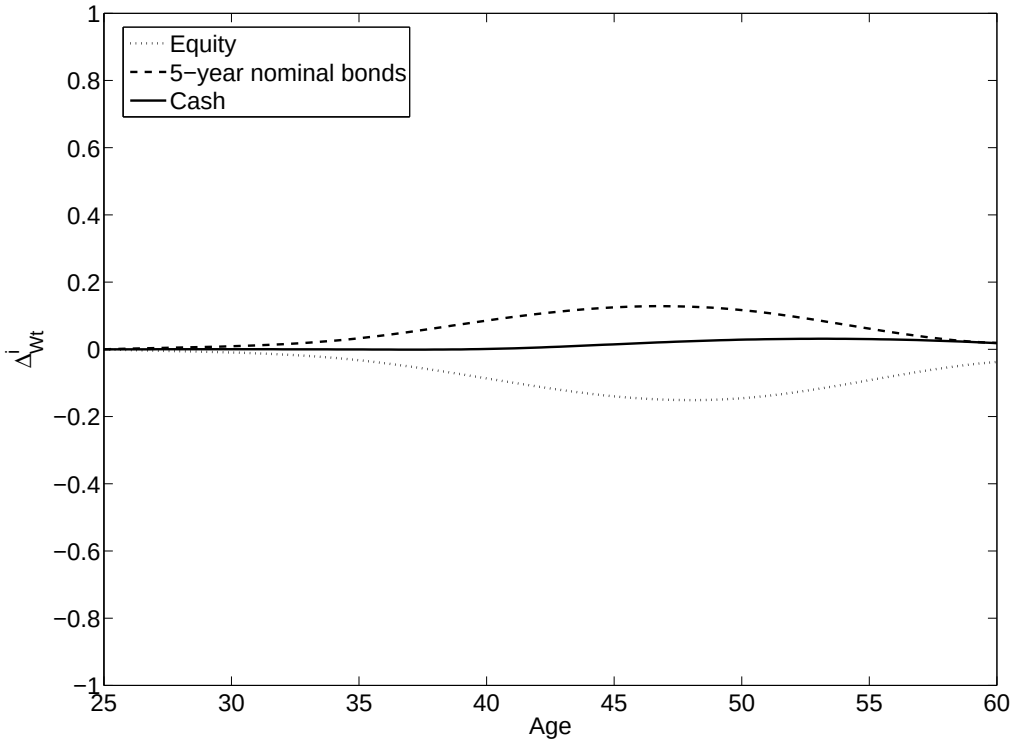


Figure 1.5: Tilts in the asset allocation induced by financial wealth

Tilts in the optimal allocation to stocks, 5-year nominal bonds, and cash in response to changes in financial wealth. We portray the difference Δ^i_{Wt}, in which i indicates the asset, in the optimal allocation in case financial wealth equals its 75% and 25% quantile. The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. The vertical axis displays the difference in the optimal allocation and the horizontal axes the individual's age. We present five-year averages to summarize the results.

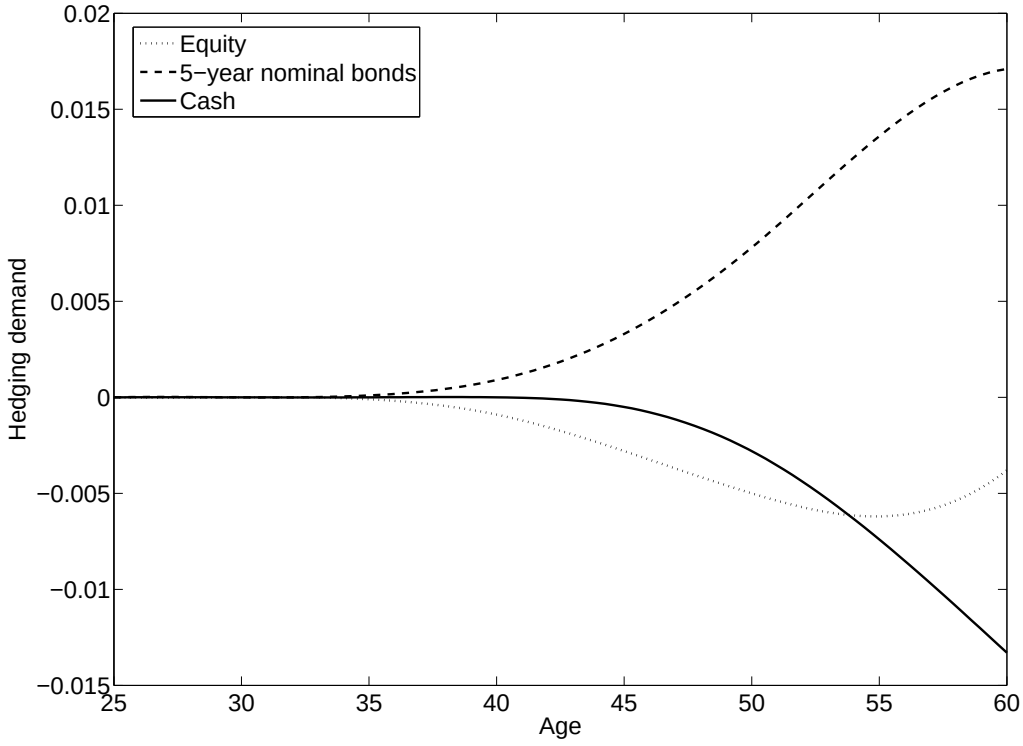


Figure 1.6: Hedging demands induced by time-varying bond risk premia over the life-cycle
Hedging demands using stocks, 5-year nominal bonds, and cash induced by time-varying bond risk premia. The hedging demands are calculated at each point in the life cycle by comparing the optimal strategies for the Strategic Investor and the Conditionally Myopic Investor. The main text provides further details. The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. The vertical axis displays the hedging demands and the horizontal axis the individual's age. We present five-year averages to summarize the results.

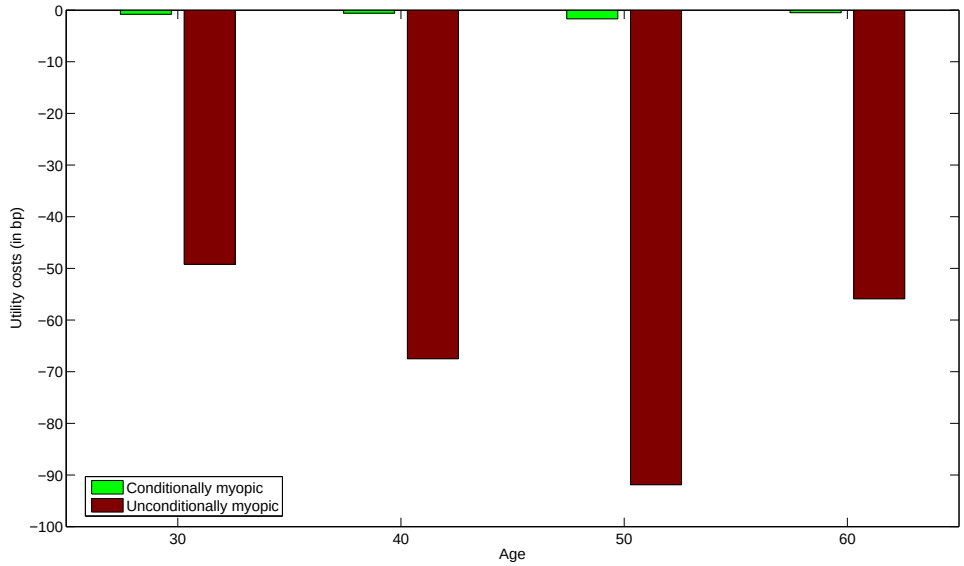


Figure 1.7: Utility costs of sub-optimal strategies over the life-cycle

Utility costs of sub-optimal strategies over the individual's life cycle. We compare the Strategic Investor, Conditionally Myopic Investor, and the Unconditionally Myopic Investor by calculating utility costs relative to the first strategy. Utility costs are determined as a fraction of certainty equivalent consumption and expressed in basis points (bp). The main text provides further details. The individual's coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$. The vertical axis displays the utility costs and the horizontal axis the individual's age.

Panel A: Tilts in the asset allocation

	$\gamma = 3$						$\gamma = 7$					
	Real rate			Expected inflation			Real rate			Expected inflation		
Age	Equity	Bonds	Cash	Equity	Bonds	Cash	Equity	Bonds	Cash	Equity	Bonds	Cash
26-30	0%	0%	0%	0%	0%	0%	0%	0%	0%	-4%	4%	0%
31-35	0%	0%	0%	-5%	5%	0%	-3%	3%	0%	-29%	29%	0%
36-40	0%	0%	0%	-7%	7%	0%	-9%	9%	0%	-49%	51%	-1%
41-45	0%	0%	0%	-13%	13%	0%	-12%	11%	1%	-55%	66%	-11%
46-50	-2%	2%	0%	-24%	24%	0%	-12%	10%	2%	-49%	73%	-24%
51-55	-6%	6%	0%	-38%	38%	0%	-10%	8%	2%	-41%	77%	-36%
56-60	-10%	9%	0%	-57%	57%	0%	-7%	6%	1%	-35%	84%	-49%
61-65	-10%	8%	1%	-57%	67%	-9%	-5%	4%	1%	-23%	90%	-67%

Panel B: Hedging demands

	$\gamma = 3$			$\gamma = 7$		
Age	Equity	Bonds	Cash	Equity	Bonds	Cash
26-30	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
31-35	0.0%	0.0%	0.0%	-0.3%	0.3%	0.0%
36-40	0.0%	0.0%	0.0%	-0.8%	0.9%	-0.2%
41-45	0.0%	0.0%	0.0%	-1.0%	1.7%	-0.7%
46-50	0.0%	0.0%	0.0%	-1.0%	2.7%	-1.7%
51-55	-0.1%	0.1%	0.0%	-0.9%	3.7%	-2.8%
56-60	-0.2%	0.2%	0.0%	-0.4%	3.9%	-3.5%

Panel C: Utility costs (in bp)

	$\gamma = 3$		$\gamma = 7$	
Age	Cond. Myopic	Unc. Myopic	Cond. Myopic	Unc. Myopic
30	0	-21	-4	-79
40	0	-35	-6	-91
50	0	-59	-7	-111
60	0	-51	-2	-75

Table 1.4: Alternative risk preferences

This table presents tilts in the optimal allocation to stocks, 5-year nominal bonds, and cash (Panel A), hedging demands (Panel B), and utility costs induced by following the optimal strategy of the Conditionally Myopic Investor (Cond. Myopic) or the Unconditionally Myopic Investor (Unc. Myopic), (Panel C). The coefficient of relative risk aversion equals either $\gamma = 3$ (left) or $\gamma = 7$ (right). The time preference parameter equals $\beta = 0.96$.

Panel A: Tilts in the asset allocation

	No High School						College					
	Real rate			Expected inflation			Real rate			Expected inflation		
Age	Equity	Bonds	Cash	Equity	Bonds	Cash	Equity	Bonds	Cash	Equity	Bonds	Cash
26-30	0%	0%	0%	-2%	2%	0%	0%	0%	0%	-1%	1%	0%
31-35	0%	0%	0%	-17%	17%	0%	0%	0%	0%	-11%	11%	0%
36-40	-5%	5%	0%	-32%	32%	0%	-1%	1%	0%	-22%	22%	0%
41-45	-10%	10%	0%	-49%	51%	-2%	-9%	8%	0%	-43%	44%	0%
46-50	-13%	12%	1%	-56%	65%	-9%	-13%	12%	1%	-56%	62%	-6%
51-55	-11%	10%	1%	-52%	71%	-18%	-11%	10%	1%	-53%	70%	-17%
56-60	-9%	8%	1%	-47%	77%	-30%	-9%	8%	1%	-48%	77%	-29%
61-65	-6%	5%	1%	-35%	81%	-47%	-6%	5%	1%	-35%	81%	-46%

Panel B: Hedging demands

	No High School			College		
Age	Equity	Bonds	Cash	Equity	Bonds	Cash
26-30	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
31-35	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
36-40	-0.2%	0.2%	0.0%	-0.1%	0.1%	0.0%
41-45	-0.4%	0.6%	-0.1%	-0.4%	0.5%	-0.1%
46-50	-0.6%	1.1%	-0.5%	-0.6%	1.0%	-0.4%
51-55	-0.6%	1.7%	-1.0%	-0.7%	1.6%	-0.9%
56-60	-0.4%	1.9%	-1.5%	-0.4%	1.9%	-1.5%

Panel C: Utility costs (in bp)

	No High School		College	
Age	Cond. Myopic	Unc. Myopic	Cond. Myopic	Unc. Myopic
30	-1	-58	-1	-61
40	-1	-82	-1	-74
50	-2	-99	-2	-97
60	-1	-58	0	-58

Table 1.5: Alternative education levels

This table presents tilts in the optimal allocation to stocks, 5-year nominal bonds, and cash (Panel A), hedging demands (Panel B), and utility costs induced by following the optimal strategy of the Conditionally Myopic Investor (Cond. Myopic) or the Unconditionally Myopic Investor (Unc. Myopic), (Panel C). The individual's education level is either "No High School" (left) or "College" (right). The coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$.

	Panel A: Tilts in the asset allocation											
	Correlation income risk and equity returns is $\rho = 25\%$						10-year nominal bonds					
		Real rate			Expected inflation			Real rate			Expected inflation	
Age	Equity	5-year bonds	Cash	Equity	5-year bonds	Cash	Equity	10-year bonds	Cash	Equity	10-year bonds	Cash
26-30	-2%	2%	0%	-33%	33%	0%	3%	-3%	0%	-44%	44%	0%
31-35	-9%	9%	0%	-63%	63%	0%	10%	-10%	0%	-68%	68%	0%
36-40	-13%	13%	0%	-72%	73%	0%	13%	-13%	0%	-72%	72%	0%
41-45	-16%	15%	1%	-78%	81%	-3%	13%	-14%	1%	-78%	79%	-1%
46-50	-17%	15%	2%	-70%	84%	-14%	10%	-14%	4%	-82%	84%	-2%
51-55	-13%	11%	2%	-58%	81%	-23%	7%	-15%	8%	-72%	86%	-13%
56-60	-9%	8%	1%	-49%	82%	-33%	4%	-17%	13%	-57%	86%	-29%
61-65	-6%	5%	1%	-35%	84%	-48%	1%	-19%	18%	-36%	83%	-47%

	Panel B: Hedging demands					
	$\rho = 25\%$			10-year nominal bonds		
Age	Equity	Bonds	Cash	Equity	Bonds	Cash
26-30	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
31-35	0.0%	0.0%	0.0%	-0.1%	0.1%	0.0%
36-40	-0.3%	0.3%	0.0%	-0.5%	0.5%	0.0%
41-45	-0.4%	0.7%	-0.2%	-1.0%	1.1%	-0.1%
46-50	-0.5%	1.3%	-0.7%	-1.4%	1.9%	-0.5%
51-55	-0.6%	1.9%	-1.3%	-1.6%	2.7%	-1.1%
56-60	-0.4%	2.1%	-1.7%	-0.9%	2.7%	-1.8%

	Panel C: Utility costs (in bp)				
	$\rho = 25\%$		10-year nominal bonds		
Age	Cond. Myopic	Unc. Myopic		Cond. Myopic	Unc. Myopic
30	-1	-75		-4	-184
40	-1	-92		-5	-257
50	-2	-99		-9	-292
60	0	-57		-2	-172

Table 1.6: Correlation between income risk and asset returns, and alternative asset menus

This table presents tilts in the optimal allocation to stocks, 5-year (left) or 10-year (right) nominal bonds, and cash (Panel A), hedging demands (Panel B), and utility costs induced by following the optimal strategy of the Conditionally Myopic Investor (Cond. Myopic) or the Unconditionally Myopic Investor (Unc. Myopic), (Panel C). The left side of the table presents the results when income risk and equity returns have a correlation of $\rho = 25\%$. The coefficient of relative risk aversion equals $\gamma = 5$ and the time preference parameter $\beta = 0.96$.

Chapter 2

Optimal Annuity Risk Management

Abstract

We study the optimal consumption and portfolio choice problem over an individual's life-cycle taking into account annuity risk at retirement. Optimally, the investor allocates wealth at retirement to nominal, inflation-linked, and variable annuities and conditions this choice on the state of the economy. We also consider the case in which there are, either for behavioral or institutional reasons, limitations in the types of annuities that are available at retirement. Subsequently, we determine how the investor optimally anticipates annuitization before retirement. We find that (i) using information on term structure variables and risk premia significantly improves the annuity choice, (ii) annuity market incompleteness is economically costly, and (iii) adjustments in the optimal investment strategy before retirement induced by the annuity demand due to inflation risk and time-varying risk premia are significant in economic terms. This latter result holds as well for sub-optimal annuity choices. The adjustment to hedge real interest rate risk is negligible.

2.1 Introduction

In an economy with complete annuity markets, fairly-priced annuities, and households with no preference to bequeath wealth to their heirs, it is optimal to transfer retirement wealth fully into annuities (Yaari (1965) and Davidoff, Brown, and Diamond (2005)). Annuities reallocate wealth from states in which the individual died, and hence derives no utility from consumption, to states in which the individual is still alive. Even if annuity markets are incomplete, it is often optimal to annuitize a sizeable part of retirement benefits, see Davidoff, Brown, and Diamond (2005). Nominal, inflation-linked, and variable annuities that are linked to a broad equity index allow households to manage exposures to interest rate, inflation, and equity risk, including the corresponding risk premia, at retirement. However, annuitization exposes the investor to annuity risk: the utility derived from the annuity payoffs may disappoint if financial market conditions turn out to be unfavorable at retirement.[1] This paper analyzes the utility gains that are realized by the optimal annuity choice at retirement

[1] Soares and Warshawsky (2004) illustrate annuity risk in value terms by determining the historical initial payout of both nominal and inflation-linked annuities.

as well as how investors optimally anticipate annuity risk before retirement. This latter question is relevant both for optimal and for sub-optimal annuity choices. In addition, we quantify the utility costs of annuity market incompleteness.

The reason why annuity risk is important to individuals is that insurance companies typically do not repurchase annuity products or at highly unfavorable prices only. Irreversibility is a direct consequence of the adverse selection problem in annuity markets as annuitants generally possess better information concerning their health status, in particular when they are in bad health. It is therefore hard, if not impossible, to dynamically rebalance the annuity portfolio in response to changes in financial markets after retirement. Acknowledging the illiquidity of annuity products after retirement, the investor can essentially manage annuity risk along two lines.[2] First, the optimal annuity portfolio at retirement can incorporate financial market conditions at this date. Second, by trading equity and bonds before retirement, the investor can construct hedging portfolios that pay off if the state of the economy is unfavorable at retirement. Obviously, the optimal investment strategy in the period before retirement depends on the annuitization strategy followed. Even if, either for behavioral or institutional reasons, investors restrict attention to only part of the annuity menu mentioned above,[3] hedging risks before retirement is welfare improving.

Before retirement, the investor can use equity and bond markets to engage in (dynamic) trading strategies to hedge annuity risk. The recent long-term asset allocation literature does not (explicitly) account for the state dependence of the value function at retirement as a result of annuity risk. Notable exceptions are Boulier, Huang, and Taillard (2001), Deelstra, Grasselli, and Koehl (2003), and Cairns, Blake, and Dowd (2006) in which the investor respectively hedges a minimal guarantee or the interest rate risk induced by an annuity-like product at retirement. These papers, however, restrict attention to a single annuity product at retirement and abstract from risks relevant for long-term investors like inflation and changes in risk premia. In fact, recent developments in the dynamic asset allocation literature emphasize the importance for long-term asset allocation of time variation in interest rates, see Brennan and Xia (2000), and Wachter (2003), inflation rates, see Campbell and Viceira (2001b) and Brennan and Xia (2002), and risk premia, see Brandt (1999), Campbell and Viceira (1999), Wachter (2002), Campbell, Chan, and Viceira (2003), Sangvinatsos and Wachter (2005), Brennan and Xia (2002), and Koijen, Nijman, and Werker (2007a). We show that the same risk factors play a role in the optimal (conditional) demand for annuities at retirement and the induced hedging strategies in the period before retirement.[4]

[2]Browne, Milevsky, and Salisbury (2003) determine a liquidity premium required by investors to compensate for the illiquidity of annuities.

[3]See for instance Diamond (1997) and Brown and Poterba (2001).

[4]Bodie and Pesando (1983) show that annuity products inherit the risk-return characteristics from the asset underlying the annuity.

We study the optimal portfolio, consumption, and annuity choice over the investor's life-cycle. Our financial market model allows for time-varying interest rates, inflation rates, and risk premia. At retirement, we determine the optimal allocation to all three annuity products, conditional upon the state of the economy, thus extending the class of annuitization strategies considered so far in the literature. In addition, we estimate the welfare loss of ignoring conditioning information or, as often observed in real-world annuity markets, restricting attention to only part of the annuity menu. We subsequently solve for the optimal investment and consumption strategy in the period before retirement. During this stage of the investor's life-cycle, the investor receives a stream of labor income of which a fixed fraction is saved for retirement purposes. The savings are allocated dynamically to stocks, nominal and inflation-linked bonds, and a nominal cash account. We derive optimal investment strategies using equity and bonds to hedge annuity risk induced by both optimal and sub-optimal annuitization strategies. We compute the welfare costs of ignoring annuity risk all together in the investment strategy in the period before retirement.

This paper is the first to provide an integral solution to the investment problems caused by annuity risk in a market in which interest rates, inflation, and risk premia are time varying, and provides therefore three main contributions to the extant literature. First, we show that conditioning the annuity choice on financial market conditions improves welfare significantly. The welfare costs of ignoring information on the term structure and risk premia at retirement range from 7%-9% of certainty equivalent consumption during retirement, depending on the investor's risk preferences. The optimal conditional annuity strategy turns out to be a complex function of the state variables, which may limit its practical use. However, we show that 75%-95% of the gains due to incorporating conditioning information can be obtained by following a simple linear portfolio rule. Second, we quantify the welfare costs of annuity market incompleteness, and find that these costs can be economically significant. Restricting access to nominal annuities is mainly costly for conservative investors, while only having access to inflation-linked annuities is most harmful for aggressive investors. For an individual with average risk aversion ($\gamma = 5$), the costs of investing retirement wealth fully in nominal annuities is estimated to be 28%, and 14% if the individual restricts attention to inflation-linked annuities. This implies that annuity market incompleteness is economically costly and corroborates the results of Brown and Poterba (2001) and Blake, Cairns, and Dowd (2003) who confine attention to a few possible retirement strategies. Third, we determine the optimal hedging strategy in the period before retirement for four different annuitization strategies. The optimal conditional annuitization strategy invests in all three annuities and the weights are state dependent. The optimal unconditional strategy also uses all annuity products, but its allocation is independent of the state of the economy at retirement. The third and fourth annuitization strategies invest all wealth accumulated in

nominal or inflation-linked annuities, respectively. We find that the (additional) welfare costs of not hedging annuity risk before retirement equal 9% for the first two annuitization strategies and 1% for nominal annuitization for an individual with an average risk aversion ($\gamma = 5$). However, the (additional) welfare costs of not hedging the annuitization strategy which invests all wealth in inflation-linked annuities is negligible. This leads to the conclusion that hedging annuity risk induced by time variation in inflation and risk premia is welfare improving, while the annuity risk caused by real interest rates is only of minor importance. The limited impact of real interest rate risk is a consequence of the relatively strong mean-reversion implied by our estimates.

Our analysis delivers several policy implications for risk management and product design of defined benefit (DB) and defined contribution (DC) pension plans. In case of DB pension plans, participants are generally entitled to nominal or inflation-linked annuities. The number of annuities to be received at retirement is linked to the participant's average or final wage. This liability induces significant annuity risk at the fund level, predominantly caused by time variation in inflation rates. In addition, we find that restricting the annuity menu to either nominal or inflation-linked annuities is costly in welfare terms. This pleads for more flexible payout options in DB pension plans. Concerning DC schemes, we show that it is possible to design simple products which largely implement the optimal conditional annuitization strategy. Likewise, given the importance of hedging annuity risk in the period before retirement, DC pension plans may design products that hedge the most important sources of annuity risk that we identify, namely inflation risk and changes in risk premia, instead of simply aiming at wealth at retirement which is then subject to interest rate and inflation risk at conversion.

Optimal annuity choice has been addressed before in Charupat and Milevsky (2002), Brown and Poterba (2001), Blake, Cairns, and Dowd (2003), and Horneff, Maurer, Mitchell, and Dus (2006). Charupat and Milevsky (2002) assume interest rates and risk premia to be constant and abstract from inflation risk. They show that the optimal allocation to fixed (i.e., nominal or inflation-linked) and variable annuities coincides with the optimal allocation to stocks and bonds in the period before retirement. Brown and Poterba (2001) are mainly interested in the welfare effects of having access to inflation-linked and variable annuities. They assume that the individual converts all retirement capital to a single annuity product, or splits wealth equally between inflation-linked and either nominal or variable annuities. Brown and Poterba (2001) find that both variable annuities and inflation-linked annuities can be welfare enhancing, depending on the risk preferences of the annuitant. Blake, Cairns, and Dowd (2003) and Horneff, Maurer, Mitchell, and Dus (2006) consider various distribution programs of retirement wealth. They consider portfolios containing equity and fixed annuities and show that the ability to invest in equities during retirement

can improve welfare significantly. However, both Brown and Poterba (2001), Blake, Cairns, and Dowd (2003), and Horneff, Maurer, Mitchell, and Dus (2006) assume risk premia to be constant and do not explore the possibility to tailor the annuity portfolio to the state of the economy at retirement. We relax both assumptions and find significant additional gains in utility terms.

We simplify the problem along various dimensions for ease of exposition. First, we assume that annuities are fairly priced, which is (from the annuitant's perspective) an attractive representation of observed annuity markets. Friedman and Warshawsky (1990) and Mitchell, Poterba, Warshawsky, and Brown (1999) show that annuity products may be expensive in fair value terms due to the adverse selection problem in the annuity market. However, Mitchell, Poterba, Warshawsky, and Brown (1999) provide evidence that the deviation from the fair value of the annuity decreased substantially during the last decade. Finkelstein and Poterba (2002) also conclude that annuities are priced fairly in the UK market. Second, the investor annuitizes wealth at a single point in time. Milevsky and Young (2003) and Neuberger (2003) show that, once annuitization is irreversible, it may be optimal to transfer retirement wealth gradually into annuities. Third, we abstract from bequest motives, health shocks, and the amount of wealth pre-annuitized, see Brown (2001), Brown and Poterba (2001), and Lopes (2005), which may affect the fraction of wealth annuitized. Finally, we abstract from joint life annuities, see for instance Brown and Poterba (2000) and Brown (2001), and consider only individual immediate annuities. We leave these extensions for future research.

This paper proceeds as follows. In Section 2.2 we provide our model of the financial market and the individual's preferences, and discuss the annuity market we consider. Next, we determine in Section 2.4 the optimal conditional and unconditional annuitization strategy at retirement. We also determine the welfare costs of sub-optimal strategies and annuity market incompleteness. In Section 2.5, we solve for the optimal policies before retirement with annuity risk induced by both optimal and sub-optimal annuitization strategies. We determine furthermore the welfare costs of not accounting for annuity risk in the period before retirement. Finally, Section 2.6 concludes. Five appendices contain technical details and proofs. All tables and figures are presented in Appendix 2.F.

2.2 Financial market, annuity market, and preferences

2.2.1 Financial market

Our financial market accommodates time variation in real interest rates, inflation rates, and risk premia. The financial market model we consider is closely related to the models of

Brennan and Xia (2002) and, in discrete time, Campbell and Viceira (2001b). These papers propose two factor models of the term structure, where the factors are identified with the real interest rate (r) and expected inflation (π). Both models assume that bond risk premia are constant. We accommodate time variation in both equity and bond risk premia. The investor's asset menu comprises stocks, nominal and inflation-linked bonds, and a nominal cash account.

We assume that the real rate is driven by a single factor, X_1,

$$r_t = \delta_r + X_{1t}, \delta_r > 0, \tag{2.1}$$

and expected inflation is affine in a second, possibly correlated, factor, X_2,

$$\pi_t = \delta_\pi + X_{2t}, \delta_\pi > 0. \tag{2.2}$$

It is well known that real interest rates and expected inflation are persistent processes. Therefore, we model both factors as Ornstein-Uhlenbeck processes, with $i = 1, 2$,

$$\mathrm{d}X_{it} = -\kappa_i X_{it}\mathrm{d}t + \sigma_i'\mathrm{d}Z_t, \kappa_i > 0, \tag{2.3}$$

in which Z denotes a five-dimensional vector of independent Brownian motions and $\sigma_i \in \mathbb{R}^5$. All correlations between the factors are captured by the volatility vectors. Realized inflation is subsequently modeled as

$$\frac{\mathrm{d}\Pi_t}{\Pi_t} = \pi_t\mathrm{d}t + \sigma_\Pi'\mathrm{d}Z_t, \tag{2.4}$$

in which Π_t denotes the level of the (commodity) price index at time t and $\sigma_\Pi \in \mathbb{R}^5$.

The value of the equity index at time t is denoted by S_t, with dynamics

$$\frac{\mathrm{d}S_t}{S_t} = \mu_t\mathrm{d}t + \sigma_S'\mathrm{d}Z_t, \tag{2.5}$$

where $\mu_t = R_t + \mu_0 + \mu_1'Y_t$, in which R_t denotes the instantaneous nominal short rate, which is derived in (2.12) below, and Y a vector of forecasting variables. Risk premia are allowed to depend on the term structure variables, (X_1, X_2), and the dividend yield, D. Ang and Bekaert (2007) show that the predictive power of the dividend yield is enhanced in a joint model with the short rate.[5] We therefore take $Y = (X_1, X_2, D)'$. In order to

[5]See Ang and Bekaert (2007), Goyal and Welch (2003), Campbell and Yogo (2006), Campbell and Thompson (2007), Lettau and van Nieuwerburgh (2006) for a recent discussion on the predictive power of the dividend yield. Binsbergen and Koijen (2007) use a present-value model to show that the price-dividend ratio has strong predictive power for both future returns and future dividend growth rates.

accommodate first order autocorrelation in the dividend yield, we model the dividend yield using an Ornstein-Uhlenbeck process

$$\mathrm{d}D_t = \kappa_D \left(\mu_D - D_t\right) \mathrm{d}t + \sigma_D' \mathrm{d}Z_t, \tag{2.6}$$

with $\sigma_D \in \mathbb{R}^5$. Without further restrictions, the volatility vectors of the different processes are statistically not identified. Therefore, we impose the volatility matrix $(\sigma_1', \sigma_2', \sigma_\Pi', \sigma_S', \sigma_D')$ to be lower triangular.

To derive the prices of both nominal and inflation-linked bonds, we assume that the prices of real interest rate and inflation risk are affine in the state variables. Formally, in the nominal state price density, ϕ, with corresponding dynamics

$$\frac{\mathrm{d}\phi_t}{\phi_t} = -R_t \mathrm{d}t - \Lambda_t' \mathrm{d}Z_t, \tag{2.7}$$

we assume the prices of risk, Λ_t, to be affine in the term structure variables and the dividend yield,

$$\Lambda_t = \Lambda_0 + \Lambda_1 Y_t. \tag{2.8}$$

Bond risk premia are allowed to be time varying, but we impose restrictions on Λ_1 such that the risk premium on inflation-linked bonds is driven only by the real rate. Similarly, the risk premium on nominal bonds depends on both the real rate and expected inflation, in line with Koijen, Nijman, and Werker (2007a). We assume, in addition, that the dividend yield does not drive the term structure of interest rates, which requires the price of unexpected inflation, $\sigma_\Pi' \Lambda_t$, to be independent of the dividend yield. More formally, these restriction imply that we parameterize Λ_1 as

$$\Lambda_1 = \begin{pmatrix} \Lambda_{1(1,1)} & 0 & 0 \\ 0 & \Lambda_{1(2,2)} & 0 \\ 0 & 0 & 0 \\ \Lambda_{1(4,1)} & \Lambda_{1(4,2)} & \Lambda_{1(4,3)} \\ 0 & 0 & \Lambda_{1(5,3)} \end{pmatrix}, \tag{2.9}$$

in which the parameters in the last two rows are identified via $\sigma_S' \Lambda_1 = \mu_1'$ and $\sigma_\Pi' \Lambda_{1(:,3)} = 0$.

Given the nominal state price density in (2.7), we find for the dynamics of the real state

price density, ϕ^R,

$$\frac{\mathrm{d}\phi_t^R}{\phi_t^R} = -(R_t - \pi_t + \sigma_\Pi' \Lambda_t)\mathrm{d}t - (\Lambda_t' - \sigma_\Pi')\mathrm{d}Z_t \tag{2.10}$$

$$= -r_t \mathrm{d}t - (\Lambda_t' - \sigma_\Pi')\mathrm{d}Z_t, \tag{2.11}$$

which implies for the instantaneous nominal short rate

$$R_t = \delta_R + \left(\iota_2' - \sigma_\Pi' \tilde{\Lambda}_1\right) X_t, \tag{2.12}$$

where $\delta_R = \delta_r + \delta_\pi - \sigma_\Pi' \Lambda_0$ and $\tilde{\Lambda}_1$ denotes the first two columns of Λ_1.[6] The conditions specified in Duffie and Kan (1996) are satisfied, implying that both nominal and real bond prices are exponentially affine in the state variables. Hence, we find for the prices of a nominal bond at time t, which matures at time $t + \tau$,

$$P(X_t, t, t+\tau) = \exp(A_\tau + B_\tau' X_t), \tag{2.13}$$

and, similarly, for an inflation-linked bond

$$P^R(X_t, t, t+\tau) = \exp(A_\tau^R + B_\tau^{R\prime} X_t), \tag{2.14}$$

where A_τ, B_τ, A_τ^R, B_τ^R, and the corresponding derivations, are provided in Appendix 2.A.

2.2.2 Annuity market

Current annuity markets provide, broadly speaking, three types of individual immediate annuity products, see also Brown et al. (2001).[7] Nominal annuities ensure a constant nominal periodic payment during the remainder of the annuitant's life. Inflation-linked annuities, on the other hand, provide payments which are constant in real terms. Consequently, inflation-linked annuities can protect the annuitant against inflation risk. The third annuity product we consider is a so-called variable annuity. The payments provided by variable annuities are linked to a broad equity index.[8] In this way, the annuitant is able to benefit from the

[6] Recall that we have assumed that the price of unexpected inflation risk does not depend on the dividend yield, i.e. $\sigma_\Pi' \Lambda_{1(:,3)} = 0$.

[7] In this paper, we confine attention to immediate annuities, which implies that the payments start the period after the annuity has been purchased. Alternatively, investors can purchase deferred annuities during the accumulation phase. These products may be particularly interesting from a tax perspective. Poterba (1997) and Blake (1999) provide a detailed overview of the different annuity products offered.

[8] A variable annuity contract in the United States refers to a tax-sheltered retirement savings plan to which periodic payments are made, see Charupat and Milevsky (2002). At retirement, the wealth accumulated can be converted into an annuity, but the investor is allowed to select the underlying portfolio composition which will determine the annuity income. For instance, Bodie and Pesando (1983)consider variable annuities which

(possibly) attractive investment opportunities offered by equity markets during the retirement phase. Using nominal, inflation-linked, and variable annuities, the annuitant is able to manage exposures to the various risk factors like interest rate risk, inflation risk, and equity risk, and the annuity market is, in essence, complete. The remainder of this section discusses the pricing of annuities as well as the income streams provided in more detail.

We price the annuity products in the before-mentioned financial market. We assume throughout that annuities are fairly priced and the proper survival probabilities are taken into account.[9] We assume in addition that longevity risk is idiosyncratic.[10] Denote the probability that the annuitant, who is currently T years old, survives at least another s years by ${}_sp_T$. We normalize the nominal and real annuity payments of respectively nominal and inflation-linked annuities to one. Formally, the nominal rate of income provided by the nominal annuity at time $T+s$ for an annuity purchased at time T, $I^N(T+s,T)$, is given by $I^N(T+s,T)=1$. For inflation-linked annuities, the nominal rate of income, $I^R(T+s,T)$, is given by $I^R(T+s,T)=\Pi_{T+s}\Pi_T^{-1}$. Consequently, the price of a nominal annuity starting at time t for an annuitant of age T is given by

$$A^N\left(X_t,T\right) \;=\; \mathbb{E}_t\left(\int_0^\infty {}_sp_T I^N(T+s,T)\frac{\phi_{t+s}}{\phi_t}\mathrm{d}s\right)=\int_0^\infty {}_sp_T P\left(X_t,t,t+s\right)\mathrm{d}s. \quad (2.15)$$

The price of an inflation-linked annuity starting at time t for an annuitant of age T equals

$$A^R\left(X_t,T\right) \;=\; \mathbb{E}_t\left(\int_0^\infty {}_sp_T I^R(T+s,T)\frac{\phi_{t+s}}{\phi_t}\mathrm{d}s\right)=\int_0^\infty {}_sp_T P^R\left(X_t,t,t+s\right)\mathrm{d}s. \quad (2.16)$$

The pricing and payout structure of variable annuities is somewhat more involved, see also Bodie and Pesando (1983) and Brown and Poterba (2001). A variable annuity is parameterized by a so-called assumed interest rate (AIR), h. The AIR is an actuarial construct to determine the number of contracts obtained per dollar invested. Formally, for every dollar invested in a variable annuity, the annuitant receives $A^V(h,T)^{-1}$ contracts, with

$$A^V(h,T)=\int_0^\infty {}_sp_T e^{-hs}\mathrm{d}s. \quad (2.17)$$

are backed by a diversified portfolio of both stocks and bonds. We define, following Brown and Poterba (2001), variable payout annuities that are exclusively linked to a broad equity index and introduce the two other available annuity contracts (nominal and inflation-linked) in turn.

[9]Despite the evidence that annuities are expensive once compared to their value due to the adverse selection problems in annuity markets, see for instance Friedman and Warshawsky (1990) and Mitchell, Poterba, Warshawsky, and Brown (1999), the latter study also provide evidence that the deviation from the fair value decreased during the last decade. Finkelstein and Poterba (2002) show in addition that annuities are priced fairly in the UK market.

[10]It is straightforward to generalize the model to allow for a systematic component in longevity risk.

The rate of income provided at time $T+s$ for a variable annuity purchased at time T is given by $I^V(h, T+s, T)$, with

$$I^V(h, T+s, T) = \frac{1}{A^V(h,T)} \frac{S_{T+s}}{S_T} e^{-hs}. \tag{2.18}$$

2.2.3 Investor's preferences and labor income

The investor is assumed to participate in the labor market during the period $[t_0, T]$ and the retirement date T is specified exogenously.[11] The nominal rate of income is denoted by $L_t^{\$}$ and its real counterpart by $L_t = L_t^{\$}\Pi_t^{-1}$. Before retirement, the investor allocates wealth dynamically to stocks, two long-term nominal bonds, and a long-term inflation-linked bond.[12] The optimal proportion of wealth allocated to these assets at time t is denoted by x_t. The remainder, $1 - x_t'\iota$, is invested in a nominal cash account. In addition, the investor optimally decides upon the amount to consume at time t, C_t. At age T the investor retires and annuitizes all wealth accumulated. The fractions allocated to the nominal, inflation-linked, and variable annuity at time T are denoted by α_T^N, α_T^R, and α_T^V, respectively.

The investor derives utility from real consumption during the life-cycle, in line with Brennan and Xia (2002) and Sangvinatsos and Wachter (2005). The preferences are represented by a time-separable CRRA utility index, i.e., the value function of the problem is

$$J\left(W_{t_0}, Y_{t_0}, L_{t_0}, t_0\right) = \max_{(C_t)_{t\in[t_0,\infty)}, (x_t)_{t\in[t_0,T]}, \alpha_T^N, \alpha_T^R, \alpha_T^V} \mathbb{E}_{t_0}\left(\int_{t_0}^{\infty} {}_{t-t_0}p_{t_0} \frac{e^{-\beta t}}{1-\gamma}\left(\frac{C_t}{\Pi_t}\right)^{1-\gamma} \mathrm{d}t\right), \tag{2.19}$$

where β denotes the subjective discount factor. We assume throughout that ${}_{T-t}p_t = 1$, for all $t \in [t_0, T]$, i.e., the investor survives up to retirement with probability one. The optimization in the period before retirement is subject to a dynamic budget constraint. Let W_t denote wealth accumulated and $L_t^{\$}$ the nominal rate of labor income at time t. The budget constraint is

$$\mathrm{d}W_t = W_t\left(x_t'\Sigma\Lambda_t + R_t\right)\mathrm{d}t + \left(L_t^{\$} - C_t\right)\mathrm{d}t + W_t x_t'\Sigma \mathrm{d}Z_t,\ t_0 \le t \le T, \tag{2.20}$$

and Σ the volatility matrix of the traded assets. During the retirement phase, the investor receives annuity income. Part of this annuity income can be saved in order to smooth consumption. We assume that the wealth accumulated is invested in a nominal cash account.

[11] The model can easily be extended to allow for several predetermined dates at which the household can convert its retirement capital into annuities. We impose annuitization at a fixed date T only for computational tractability.

[12] Any additional bond is redundant as the term structure of interest rates is driven by two factors, i.e. (X_1, X_2). In order to complete the market, one inflation-linked bond is required to hedge unexpected inflation.

This leads to the budget constraint during retirement

$$\mathrm{d}W_t = W_t R_t \mathrm{d}t + (Y_t - C_t)\,\mathrm{d}t,\ t \geq T, \tag{2.21}$$

with Y_t indicating nominal annuity income at time t. Further, we assume that the investor annuitizes fully at retirement,[13] i.e., for $t \geq T$,

$$\frac{Y_t}{W_{T-}} = \frac{\alpha_T^N}{A^N\left(X_{1T}, X_{2T}, T\right)} + \frac{\alpha_T^R}{A^R\left(X_{1T}, X_{2T}, T\right)}\frac{\Pi_t}{\Pi_T} + \alpha_T^V I^V(h,t,T), \text{ with } \alpha_T^N + \alpha_T^R + \alpha_T^V = 1 \tag{2.22}$$

The budget constraint in (2.21) is subject to the initial condition $W_T = 0$, since the investor converts all wealth into annuities. Note that W_{T-} in (2.22) refers to retirement wealth just *prior to* conversion. In summary, the investor annuitizes fully at retirement and can smooth annuity income optimally during retirement using a nominal cash account, following for instance Brown and Poterba (2001).

The optimization is subject to the (institutional) constraint that annuities cannot be shorted, i.e.,

$$\alpha_T^N,\ \alpha_T^R,\ \alpha_T^V \geq 0. \tag{2.23}$$

We assume that the investor cannot capitalize future annuity income to increase today's consumption. Therefore, we impose that the investor is liquidity constrained, which formally implies, for $t > T$,

$$W_t \geq 0. \tag{2.24}$$

The dynamics of real labor income, $L_t = L_t^{\$}\Pi_t^{-1}$, are given by, with $t \in [t_0, T]$ indicating the investor's age

$$\frac{\mathrm{d}L_t}{L_t} = g_t \mathrm{d}t, \tag{2.25}$$

where g_t is calibrated on the basis of Cocco, Gomes, and Maenhout (2005) and Munk and Sørensen (2005) to capture the hump-shaped pattern in labor income over the life-cycle.

The investor's problem can be decomposed conveniently as

$$J\left(W_{t_0}, Y_{t_0}, L_{t_0}, t_0\right) = \max_{(C_t, x_t)_{t\in[t_0,T]}} \mathbb{E}_{t_0}\left(\int_{t_0}^{T} \frac{e^{-\beta t}}{1-\gamma}\left(\frac{C_t}{\Pi_t}\right)^{1-\gamma} \mathrm{d}t\right) + e^{-\beta T}\mathbb{E}_{t_0}\left(J\left(W_{T-}^R, Y_T, 0, T\right)\right) \tag{2.26}$$

[13] This assumption is made for computational tractability.

which disentangles the problem before and after retirement. We define $W^R_{T-} = W_{T-}\Pi^{-1}_{T-}$ to denote real wealth just before retirement. For future reference, we formulate after retirement the problem as, for $\gamma > 1$,

$$\begin{aligned} J\left(W^R_{T-}, Y_T, 0, T\right) &= \max_{(C_t)_{t\in(T,\infty)},\alpha^N_T,\alpha^R_T,\alpha^V_T} \mathbb{E}_T\left(\int_T^\infty {}_{t-T}p_T \frac{e^{-\beta(t-T)}}{1-\gamma}\left(\frac{C_t}{\Pi_t}\right)^{1-\gamma} \mathrm{d}t\right) \\ &= \frac{1}{1-\gamma}\left(W^R_{T-}\right)^{1-\gamma} \min_{\left(C^R_t\right)_{t\in(T,\infty)},\alpha^N_T,\alpha^R_T,\alpha^V_T} \mathbb{E}_T\left(\int_T^\infty {}_{t-T}p_T e^{-\beta(t-T)}\left(C^R_t\right)^{1-\gamma} \mathrm{d}t\right), \end{aligned} \tag{2.27}$$

with $C^R_t = C_t W^{-1}_{T-}\Pi_{T-}\Pi^{-1}_t$, i.e., C^R_t denotes the real fraction of wealth consumed at time t.

2.3 Model estimation and calibration

2.3.1 Estimation of the financial market model

The financial market model is estimated using monthly US data on bond yields, inflation, and stock returns over the period January 1952 up to May 2002. The government yield data up to February 1993 is the McCulloch and Kwon data and we extend the sample using data provided by Rob Bliss.[14] We use 3-month, 6-month, 1-year, 2-year, 5-year, and 10-year nominal yields in estimation. Inflation data is obtained via CRSP and is based on the CPI-U price index. Finally, we use returns on the CRSP value-weighted NYSE/Amex/Nasdaq index for stock returns. We construct the dividend yield on the basis of the index with and without dividends, along the lines of Campbell, Chan, and Viceira (2003). The model is estimated by means of maximum likelihood using standard Gaussian Kalman filtering techniques as detailed in Appendix 2.B. The estimation results are summarized in Table 2.1.

We find that expected inflation is far more persistent than the real interest rate (see also Brennan and Xia (2002) and Campbell and Viceira (2001b)). The innovations to the real rate and expected inflation are negatively correlated. For the unconditional prices of real interest rate risk and expected inflation risk, Λ_0, we find that the former is rewarded a much higher price of risk than the latter. This implies that real interest rate risk is a highly priced risk factor in comparison to the expected inflation factor. The unconditional equity risk premium is estimated to be 4.3%. The unconditional nominal bond risk premium equals 1.8% for a 10-year nominal bond. Likewise, the risk premium on 10-year inflation-linked bonds equals 1.2%. Concerning the dynamics of bond risk premia, we find that the real bond risk premium increases with the real rate and the nominal bond risk premium increases with both the real rate and expected inflation. We define the inflation risk premium to be the difference in

[14] We are grateful to Rob Bliss for providing the data.

expected returns on nominal and inflation-linked bonds with the same maturity, following Campbell and Viceira (2001b). As a result of the negative correlation between the real rate and expected inflation, the exposure of nominal bonds to the real rate will be smaller than for inflation-linked bonds, with the same maturity. This implies that the inflation risk premium is decreasing in the real rate and increasing in expected inflation, in line with Buraschi and Jiltsov (2005) and Koijen, Nijman, and Werker (2007a). Next, we find that the equity risk premium decreases with the real rate ($\mu_{1(1)} < 0$) and expected inflation ($\mu_{1(2)} < 0$) and is increasing in the dividend yield ($\mu_{1(3)} > 0$). Finally, the estimates of the dividend yield process reveal the well-documented persistence of financial ratios and the high (negative) correlation of its innovations with equity returns.

2.3.2 Calibration of the annuity market

Bodie and Pesando (1983) select the AIR equal to the expected return on the portfolio underlying the variable annuity. Brown and Poterba (2001) remark that commonly observed values of the AIR are around 3-4%. The choice of the AIR is not unimportant as it affects the distributional properties of the income stream offered by variable annuities. For instance, if the AIR is chosen relatively high, the number of contracts obtained will be high as well. This implies in turn that the initial payments will be high, but the income stream is expected to decline rapidly. Likewise, for low values of the AIR, the initial payout is low, but increases in expectation. Appendix 2.C provides a rigorous discussion of the role of the AIR in a model with constant investment opportunities. Throughout the paper, we consider variable annuities with an AIR of $h = 4\%$. Finally, in order to calibrate the survival probabilities, we use mortality rates observed in 1999 provided by the human mortality database.[15]

Figure 2.1 portrays the mean and volatility of the real (monthly) annuity income provided by nominal, inflation-linked, and variable annuities if the investor converts \$100,000 at retirement. The horizontal axis portrays the investor's age and the vertical axis the corresponding annuity income. The vector of state variables is set equal to its unconditional expectation. Inflation-linked annuities provide a riskless payoff stream, but the level is considerably lower than the mean payoff of variable annuities for all ages and that of nominal annuities up to age 75. The real payoffs provided by nominal and variable annuities are risky and the volatility of the payoffs increases with the investor's age.

Figure 2.2 displays the average, real (monthly) annuity income provided by nominal, inflation-linked, and variable annuities for different initial values of the real rate, expected inflation, and the dividend yield if the investor converts \$100,000 at retirement. Different values of the real rate (Panel A and Panel B) have a small effect on the level of the payoffs and

[15] We refer to `http://www.mortality.org/` for further information.

the main patterns are unaffected. This is markedly different for expected inflation. When initial expected inflation is low (Panel C), the initial payoff of nominal annuities is low and the expected payoff stream is more stable in real terms. When initial expected inflation is high (Panel D), the initial payoffs of nominal annuities are high, but decline rapidly in expectation. The opposite occurs for variable annuities, since the equity risk premium is negatively related to the level of expected inflation. The level of expected inflation is therefore likely to impact the investor's annuity choice and utility derived from inflation-sensitive annuity products. Finally, Panel E and Panel F portray the impact of different initial values of the dividend yield. Different values of the dividend yield have a substantial impact on the expected annuity payoffs of variable annuities, which is a consequence of its high persistence.

2.3.3 Calibration of labor income and preferences

The (deterministic) growth rate of labor income is given by

$$g_t = 0.1682 - 0.00646t + 0.00006t^2, \tag{2.28}$$

which corresponds to an individual with high school education in the estimates of Cocco, Gomes, and Maenhout (2005). The income rate at age t_0 is normalized to $L_{t_0} = 1$ and initial wealth $W_{t_0} = 0$. We assume that the investor does not receive any additional form of income during retirement, i.e. $L_t = 0$ for $t \geq T$. We are before retirement predominantly concerned with the impact of the annuitization decision at retirement on the optimal investment strategies before retirement. We therefore abstract from idiosyncratic labor income risk. Abstracting from idiosyncratic labor income risk allows us to solve the model in closed-form. We refer to Viceira (2001), Cocco, Gomes, and Maenhout (2005), Munk and Sørensen (2005), and Koijen, Nijman, and Werker (2007a) for an analysis of the impact of idiosyncratic labor income on optimal portfolio choice. Note that the labor income stream is equivalent to a particular portfolio of inflation-linked bonds in our model.

In terms of preference parameters, we consider three values for the coefficient of relative risk aversion, namely $\gamma = 2$, 5, and 10. The time preference parameter is set, in accordance with Cocco, Gomes, and Maenhout (2005), to $\beta = 0.04$.

2.4 Optimal retirement choice

2.4.1 Optimal annuity choice

In this section, we first determine the optimal annuity choice at retirement. Optimally, the investor's annuity menu contains nominal, inflation-linked, and variable annuities with an assumed interest rate (AIR) of $h = 4\%$. This menu allows the investor to construct the proper exposures to all risk factors in the economy. The investor determines the optimal allocation conditional on the current state of the economy. Next, we assess the welfare cost of sub-optimal annuitization strategies, like ignoring conditioning information or restricting the annuity menu to nominal or inflation-linked annuities alone. These latter annuitization strategies are particularly interesting for at least two reasons. First, individual investors tend to restrict attention, either for behavioral or institutional reasons, to nominal annuities, see Diamond (1997) and Brown and Poterba (2001). Second, Defined Benefit (DB) pension plans generally offer their participants a nominal or an inflation-linked annuity. Therefore, it is important to quantify the welfare costs of annuity market incompleteness, as well as its implications for the optimal investment strategies before retirement. This latter question is addressed in Section 2.5.

We now briefly outline the optimization problem faced by the investor during retirement as described in Section 2.2.3. The investor has to decide upon the annuity choice at retirement and, given the annuitization strategy selected at retirement, the consumption strategy during the retirement phase. The annuitization strategy optimally takes into account the amount of capital accumulated and the economic conditions at retirement as summarized by the real rate, expected inflation, and the dividend yield. During the retirement phase, the consumption-savings decision incorporates the same state variables, but in addition to these also the current income level provided by nominal and variable annuities. We use the homogeneity property of the power utility index to normalize the current wealth level to one, which reduces the number of state variables by one. Despite this simplification, the problem at retirement is characterized by three state variables at retirement and five state variables during retirement, which is computationally very hard to solve using standard dynamic programming tools. We therefore opt for a simulation-based approach which is essentially an extension of the methods developed in Brandt, Goyal, Santa-Clara, and Stroud (2005). First, we simulate 10,000 trajectories of the state variables and corresponding annuity income. Subsequently, we proceed as in case of numerical dynamic programming, but we approximate the conditional expectations we encounter by expansions in a set of basis functions in the state variables. This approach is virtually insensitive to the number of state variables and, therefore, suited to solve the problem at hand. By proceeding backwards,

we obtain an estimate of the value function at retirement for a certain retirement state and annuity portfolio. This allows us to optimize over the annuity portfolio at retirement. We optimize over a grid with step sizes of 5%. Our numerical procedure then results in a set of optimal annuity choices for all of the 10,000 initial states of the economy at retirement. A detailed discussion of the approach is provided in Appendix 2.D.

It turns out that the optimal annuitization strategy, indicated by $\alpha_T^i(Y)$, with $i \in \{N, R, V\}$,[16] is a non-linear function of the state variables at retirement (the real rate, expected inflation, and the dividend yield). In order to summarize the results, we consider a first-order approximation of the optimal annuitization strategy in these state variables. We will show in the next section that this approximation is very accurate. More specifically, we run (cross-sectional) regressions for the optimal weights in each of the annuity products on the state variables (real rate, expected inflation, and the dividend yield). We standardize the state variables to enhance the interpretation of the coefficients. Thus, we have

$$\alpha_T^i(Y) \backsimeq \alpha_0^i + \sum_{j=1}^{3} \alpha_j^i \tilde{Y}_j, \tag{2.29}$$

with $i \in \{N, R, V\}$ and $\tilde{Y} = (Y - \mathbb{E}(Y))/\sigma(Y)$. The constant term (α_0^i) can then be interpreted as the unconditional allocation to a particular annuity product. The slope coefficients (α_j^i) present the percentage change in the allocation for a one (unconditional) standard deviation increase in the j-th state variable. As all retirement wealth is annuitized at retirement, we have by construction that the sum (over the three annuity products) of the constant terms equals one, i.e., $\sum_{i=1}^{3} \alpha_0^i = 1$, and the sum of the slope coefficients for a particular state variable is zero, i.e., for $j \in \{1, 2, 3\}$, $\sum_{i=1}^{3} \alpha_j^i = 0$. These regressions therefore provide a concise interpretation of the resulting optimal annuity strategy, and how it depends on the state of the economy.

Table 2.2 presents the optimal annuity strategy at retirement. Recall that the constants indicate the unconditional allocation to a particular annuity product. For instance, an investor with a coefficient of relative risk aversion equal to $\gamma = 5$ allocates 50% to inflation-linked annuities, 42% to variable annuities, and 8% to nominal annuities. In all cases, the unconditional allocation to nominal annuities is marginal. Since expected inflation is a persistent and relatively low priced factor, risk-averse investors are not willing to bear inflation risk. Similarly, more aggressive investors consider variable annuities, which are linked to an equity index, as more attractive from a risk-return perspective. This explains the minor role of nominal annuities in the optimal retirement choice. Table 2.2 also shows that the optimal annuity choice is sensitive to the state of the economy at retirement. In particular

[16] Recall that 'N' refers to nominal, 'R' to inflation-linked or real, and 'V' to variable annuities.

changes in expected inflation and dividend yield, which are both highly persistent, can alter the optimal allocation substantially. For an investor with a risk aversion of $\gamma = 5$, the optimal allocation to variable annuities reduces by 11% in absolute terms if expected inflation increases with one (unconditional) standard deviation. In response, the allocation to nominal and variable annuities increases. Recall that nominal bond risk premia are increasing in expected inflation, whereas the equity risk premium relates negatively to expected inflation. Likewise, a one (unconditional) standard deviation increase in the dividend yield leads to a 28% increase in the allocation to variable annuities at the cost of the allocation to nominal and especially inflation-linked annuities.

Finally, we determine the optimal annuity choice for individuals who do not incorporate conditioning information in their portfolio choice. More specifically, the optimal annuity choice is determined using the unconditional value function as opposed to the conditional value function. The results are presented in Table 2.3. The optimal unconditional annuitization strategy is comparable to the optimal conditional annuitization strategy, when the state variables equal their unconditional expectation.

2.4.2 Welfare costs of sub-optimal annuitization strategies

With the optimal annuity strategies in hand, we can compute the welfare costs induced by adopting one of the four sub-optimal annuitization strategies described above. First, we consider the unconditional annuity choice, where wealth is allocated to all three annuity products, but independently of the state of the economy. Second, the annuity choice may be restricted to either nominal or inflation-linked annuities only, which quantifies the costs of annuity market incompleteness. Third, we investigate whether it is possible to approximate the optimal conditional annuity choice with a simple linear portfolio rule.

The welfare costs are displayed in Table 2.4. The optimal conditional annuity choice serves as the benchmark strategy. It turns out that taking into account information about the term structure and risk premia at retirement is significantly welfare improving. The welfare costs range from 7% up to almost 9% of certainty equivalent consumption, depending on the investor's risk preferences. In other words, investors are willing to give up 7-9% of their retirement wealth in order to optimally incorporate the economic conditions at retirement. If the retirement choice is restricted to inflation-linked annuities, we find that even conservative investors incur welfare costs close to 10%. Restricting attention to only nominal annuities induces large welfare costs, especially for conservative investors. In this case, the decrease in certainty equivalent retirement consumption ranges from 22% for aggressive investors to 55% for conservative investors. Converting retirement wealth to nominal annuities exposes the individual to inflation risk, which turns out to induce significant utility costs. These results

imply that both equity exposure during retirement and the possibility to insure inflation risk are valuable for individuals. While these results are in line with Brown and Poterba (2001) and Blake, Cairns, and Dowd (2003), recall that these studies assume risk premia to be constant and do not explore the possibility to tailor the annuity portfolio to the state of the economy at retirement.

The optimal conditional annuity choice may be a complex function of the state variables. It is therefore useful to investigate whether a simple first-order approximation to the optimal annuity weights as in (2.29) can recuperate most of the costs induced by the unconditional strategy. More precisely, we consider

$$\alpha_T^{\text{Linear},i}(Y) = \frac{\max(\alpha_T^i(Y), 0)}{\sum_{i=1}^3 \max(\alpha_T^i(Y), 0)}, \tag{2.30}$$

with α_T^i given in (2.29). The results are presented in the final row of Table 2.4. It turns out that this simple rule reduces the welfare loss relative to the unconditional annuity choice by 75% to 95%, depending on the risk attitude of the investor, to 0.3-2.1%. This illustrates that it is possible to design simple rules that are close substitutes for the optimal, non-linear, strategy.

2.5 Optimal policies before retirement

2.5.1 The optimal investment and consumption strategy

The main conclusion of Section 2.4 is that the optimal annuity strategy depends on the economic conditions at retirement and that it is costly in welfare terms to ignore this. Also, we show that annuity market incompleteness gives rise to significant utility costs. Even if the optimal annuity choice does not depend on the state of the economy, for instance if the investor restricts attention to nominal or inflation-linked annuities only, the value function at retirement does. As a result, an investor can anticipate this dependence before retirement and use financial markets to hedge exposures to interest rates, inflation rates, and risk premia. In this section, we derive the optimal investment and consumption strategy in the period before retirement, taking into account the annuity strategy at retirement. The existing literature either abstracts from the investor's desire to annuitize wealth, see among others Viceira (2001), Cocco, Gomes, and Maenhout (2005), and Koijen, Nijman, and Werker (2007a), or assumes that the investor restricts attention to a single annuity product, like in Cairns, Blake, and Dowd (2006).

The optimal investment and consumption policies are the solution to (2.26), subject to the dynamic budget constraint given in (2.20). We assume that the investor can allocate

wealth to stocks, two nominal bonds, and an inflation-linked bond. The maturities of the two nominal bonds are assumed to be three and ten years and the maturity of the inflation-linked bonds is set to ten years. The remainder is allocated to a nominal cash account. It is well known, see for instance Wachter (2002) and Liu (2007), that this optimal consumption problem cannot be solved explicitly in incomplete markets. For analytical tractability, we therefore assume that the investor consumes a fixed fraction of labor income, θ, and saves the remainder, $1-\theta$, before retirement.[17] We determine the optimal investment strategy and savings rate jointly. For the remainder of this section, it will turn out to be useful to define the real present value of future savings until retirement, for $t \in [t_0, T]$,

$$H(L_t, Y_t, \theta, t, T) = \int_t^T (1-\theta) L_s P^R(X_t, t, s) \mathrm{d}s. \tag{2.31}$$

Appendix 2.E derives the optimal solution to the investment problem before retirement. The optimal investment strategy at time t is given by

$$\begin{aligned} x_t^\star &= \left(\frac{W_t^R + H_t}{W_t^R}\right) \frac{1}{\gamma} \left(\Sigma\Sigma'\right)^{-1} \Sigma\Lambda_t + \\ &\quad \left(\frac{W_t^R + H_t}{W_t^R}\right) \left(1 - \frac{1}{\gamma}\right) \left(\Sigma\Sigma'\right)^{-1} \Sigma \left(\Sigma_Y' \left(\xi_1 + \xi_2 Y_t\right) + \sigma_\Pi\right) + \\ &\quad \left(\frac{W_t^R + H_t}{W_t^R}\right) \frac{1}{\gamma} \left(\Sigma\Sigma'\right)^{-1} \Sigma\Sigma_Y' \left(\Gamma_1(\tau) + \frac{1}{2}\left(\Gamma_2(\tau) + \Gamma_2(\tau)'\right) Y_t\right) - \\ &\quad \left(\Sigma\Sigma'\right)^{-1} \Sigma\Sigma_Y' \frac{H_{Yt}}{W_t^R} - \left(\frac{H_t}{W_t^R}\right) \left(\Sigma\Sigma'\right)^{-1} \Sigma\sigma_\Pi, \end{aligned} \tag{2.32}$$

where Σ and Σ_Y are the volatility matrices of the traded assets and the state variables, respectively, and ξ_1, ξ_2, $\Gamma_1(\tau)$ and $\Gamma_2(\tau)$ are given in Appendix 2.E, $\tau = T - t$, and H_Y denotes the partial derivative of H with respect to Y. It is important to note though that when the investor aims for real wealth at retirement rather than real wealth which is in turn converted to annuities, we have $\xi_1 = 0$ and $\xi_2 = 0$. Intuitively, ξ_1 and ξ_2 are the exposures to the risk factors of the value function at retirement, if the investor annuitizes her retirement wealth. The optimal portfolio in (2.32) contains four components which allow for an intuitive interpretation. The first component is the standard myopic demand, which exploits the risk-return trade-off provided by the assets. The second component constitutes the minimum variance portfolio in an investment problem in which $K(Y_T)$ serves as the numéraire. The third component hedges changes in the state variables which affect future investment opportunity sets. This component depends on the investment horizon. The first three components are pre-multiplied by the ratio of current real financial wealth (W_t^R) plus

[17] This setup therefore mirrors a pure DC system.

the real present value of future savings (H_t) to current real financial wealth. This tilt of the optimal portfolio is a consequence of the fact that the investor has to implement the optimal strategy via financial wealth rather than total wealth, see also Bodie, Merton, and Samuelson (1992) and Munk and Sørensen (2005). The relative importance of the first three components is determined by the investor's risk attitude. The fourth and final component (the last two terms in (2.32)) corrects the optimal investment strategy for the exposures to the risk factors of the savings stream. The optimal savings rate is determined in Appendix 2.E.

The annuitization strategy affects the optimal investment strategy before retirement via two channels. First, the minimum variance portfolio (the second component in (2.32)) depends directly on ξ_1 and ξ_2, which both are zero if the investor refrains from annuitization. Second, the coefficients $\Gamma_1(\tau)$ and $\Gamma_2(\tau)$ in the hedging portfolio (the third component in (2.32)) depend on ξ_1 and ξ_2, see Appendix 2.E. As such, the second and third component of the optimal investment strategy anticipate the investor's retirement strategy.

The optimal investment strategy in (2.32) shows that the investor's optimal portfolio is composed by three general funds and an investor-specific fund, namely the fourth component. The weights of the different components is investor-specific and depends on the income path as well as on the risk attitude of the investor.

In Section 2.5.2 we examine the adjustments in the optimal investment strategy before retirement as a result of annuity risk for both optimal and sub-optimal annuitization strategies. Subsequently, we determine in Section 2.5.3 the utility costs of sub-optimal investment strategies which abstract from the investor's preference to annuitize wealth at retirement.

2.5.2 Optimal investment and consumption with annuity risk

We now assess empirically the properties of the optimal portfolio strategy before retirement. We focus in particular on the additional hedging demands caused by annuity risk. We first of all determine the optimal investment strategy if the individual ignores annuity risk. Subsequently, we provide the adjustments to the optimal portfolio if the individual allocates retirement wealth to either the optimal conditional annuitization strategy, the optimal unconditional annuitization strategy, inflation-linked annuities, or nominal annuities.

Figure 2.3 portrays the optimal allocation to 3-year and 10-year nominal bonds, 10-year inflation-linked bonds, and stocks in absence of annuity risk. The remainder is invested in the nominal cash account. The optimal investment strategy at time t as given in (2.32) depends on the state variables (Y_t), the investor's horizon ($T-t$), the income to wealth ratio (L_t/W_t^R), and the investor's risk attitude (γ). In all graphs, the state variables are set to

their unconditional expectation, the investment horizon (i.e., time to retirement) ranges from zero to thirty years, and the income to wealth ratio equals 0.01, 0.25, 0.50, or 1. An income to wealth ratio of 0.5 implies that the investor accumulated an amount financial wealth equal to two annual incomes. In the first case ($L_t/W_t^R = 0.01$), the impact of human capital on the optimal portfolio is negligible. This allows us to separate the horizon effects generated by depletion of human capital and hedging demands as a consequence of time variation in investment opportunities. The investor's coefficient of relative risk aversion is set to $\gamma = 5$.[18] The optimal savings rate used is determined by assuming that the investor implements the optimal conditional annuitization strategy and starts saving as of age 35.

We find that the investor holds a long position in 3-year nominal bonds, which is financed by a short position in cash and 10-year nominal bonds, in line with Sangvinatsos and Wachter (2005). The role of inflation-linked bonds is limited to hedging unexpected inflation risk. Both nominal bonds and stocks exhibit strong horizon effects due to time variation in investment opportunities (i.e., $L_t/W_t = 0.01$), see also Sangvinatsos and Wachter (2005). The horizon effects are further amplified by the savings stream to which the investor is entitled. Note further that the optimal portfolio entails rather large long and short positions, in particular for nominal bonds. This is a consequence of the high correlation between the returns on these assets. Since all that matters are the induced exposures to the risk factors, the positions are in fact in economic terms less extreme than suggested by the portfolio weights. Rather than physically taking extreme bond positions, these exposures can also be created through the use of swaps or other interest rate and inflation-linked derivatives.

Next, we illustrate the impact of the annuity choice on the investor's investment strategy before retirement for the same configuration of parameters as before. In determining the hedging strategy before retirement of a particular annuitization strategy, we set the savings rate in accordance with the particular annuity choice made at retirement. The income-to-wealth ratio equals 0.5.[19] Figure 2.4 portrays the hedging strategies for, respectively, the optimal conditional annuitization strategy (Panel A), the optimal unconditional annuitization strategy (Panel B), and either only inflation-linked (Panel C) or only nominal annuities (Panel D). The hedging strategy is defined as the difference between the optimal strategy which does account for annuity risk and the strategy which ignores it. Hence, by adding the portfolios in Figure 2.3 to the hedging portfolio, we obtain the total optimal portfolio in the presence of annuity risk. First, Panel A of Figure 2.4 presents the results for the optimal conditional annuitization strategy. The hedging strategy holds long positions in 3-year nom-

[18] To conserve space, we restrict attention to an investor with coefficient of relative risk aversion equal to $\gamma = 5$. Results for any other configuration of the parameters are available upon request.

[19] The results hardly change for a reasonable range of income-to-wealth ratios. Additional results are available upon request.

inal bonds and stocks, while 10-year nominal bonds and cash are shorted. The allocation to inflation-linked bonds is hardly affected. The long-short position in nominal bonds pays off in states of the economy where either expected inflation is high or real interest rates are low. In addition, when the investor anticipates the preference to annuitize wealth at retirement, the allocation to equities increases. An important determinant of the investment opportunity set at retirement is the level of the dividend yield. Low levels of the dividend yield correspond to low expected returns on variable annuities. Given the negative correlation between innovations in equity returns and dividend yields, a long equity position in the period before retirement can hedge adverse changes of the dividend yield. Interestingly, the additional hedging demands are already substantial during early stages of the investor's life-cycle. Panel B of Figure 2.4 portrays the optimal hedging strategy for an individual implementing the optimal unconditional hedging strategy at retirement. The optimal hedging strategy of the optimal conditional and unconditional annuitization strategy are close. Next, Panel C displays the optimal hedging strategy when the investor allocates all wealth at retirement to inflation-linked annuities. The hedging portfolio is strikingly different from the one where the full annuity menu is at the individual's disposal. First, neither inflation-linked nor stocks are present in the hedging strategy. When the investor allocates all capital at retirement to inflation-linked annuities, all that matters is the exposure to the real interest rate, which is managed by positions in the 3-year and 10-year nominal bond. Second, the allocation to both nominal bonds changes. Note that the hedging strategy has only significant weights as of, say, age 60. This is a consequence of the relatively low persistence of real interest rates in comparison with expected inflation and the dividend yield implied by our estimates. This suggests that hedging inflation-linked annuities, i.e., hedging real interest rate risk, is far less important than hedging the optimal (un)conditional annuitization strategy which (possibly) allocates wealth to all three annuity products. Finally, Panel D portrays the optimal hedging strategy if the investor allocates all wealth at retirement to nominal annuities. At early stages of the investor's life-cycle, it turns out to be optimal to hold long 10-year nominal bonds, and short 3-year nominal bonds to hedge time-varying bond risk premia. Close to retirement, hedging real rate risk is more important and the hedging strategy is long in 3-year nominal bonds and shorts 10-year nominal bonds, as is the case for all other annuitization strategies.

2.5.3 Welfare costs of not hedging annuity risk

We now calculate the welfare costs of not hedging annuity risk. As before, the welfare metric employed is the amount of retirement wealth an investor is willing to give up in order to implement the optimal investment strategy before retirement. We refer to Appendix 2.E for

further details. For both investment strategies, the savings rate is determined on the basis of the investment strategy which accounts for annuity risk.

The welfare costs are presented in Table 2.5. First, we find that hedging the optimal conditional annuitization strategy, the unconditional annuitization strategy, and nominal annuities before retirement is valuable, in particular for conservative investors. Individuals with an average risk aversion ($\gamma = 5$) are willing to give up 9% of retirement consumption to hedge annuity risk induced by the optimal conditional or unconditional strategy. The welfare costs of not hedging annuity risk induced by nominal annuities amounts to 1% of retirement consumption. We find that the costs of not hedging full inflation-linked annuitization are small. This is consistent with the small additional hedging demands generated by this strategy in Figure 2.4. The bottom line is that hedging annuity risk before retirement is welfare improving due to persistent expected inflation and time variation in risk premia. The welfare costs of not hedging the annuity risk caused by inflation-linked annuities is marginal. The fact that hedging real interest rate risk is of minor importance is caused by the relatively low persistence of the real interest rate.

The welfare costs of not hedging annuity risk before retirement have been calculated for the situation in which the state variables equal their unconditional expectation. However, the costs of not hedging annuity risk are obviously dependent on the prevailing term structure and risk premia. To highlight the impact of changing state variables, we consider the case where all state variables equal their unconditional expectation up to age 55, and subsequently one of the state variables is perturbed by a one (unconditional) standard deviation, positive shock. After the shock, the state variables equal their expected value, conditional on the shock at age 55. The results are portrayed in Figure 2.5 for all four annuitization strategies.

Panel A depicts the results for the optimal conditional annuitization strategy. A positive shock at age 55 to expected inflation, leading to a lower expected returns on equity and a higher expected return on long-term nominal bonds, decreases the welfare costs of not hedging annuity risk. For a positive shock to the dividend yield, it becomes more costly for the individual to abstract from hedging before retirement. This is in line with, for instance, Brandt, Goyal, Santa-Clara, and Stroud (2005) who illustrate that the economic value of hedging demands is higher in good economic environments. This immediately implies that in this stage of the investor's life cycle, high expected inflation corresponds to a deteriorated investment opportunity set. Panel B presents the results for the optimal unconditional hedging strategy. Note that the welfare costs hardly change in response to a shock in the state variables. This implies that whenever the individual is not able to tailor the annuitization strategy to the economic conditions at retirement, the welfare costs are almost independent of the state of the economy before retirement. This is also the case for the annuity strategies

which convert all capital at retirement into inflation-linked annuities (Panel C) or nominal annuities (Panel D).

2.6 Conclusions

Households prefer to annuitize wealth fully at retirement if annuity markets are sufficiently complete, annuities are fairly priced, and if they have no bequest motive. However, this exposes investors to, what is known as, annuity risk. The utility derived from annuitized wealth may disappoint if market conditions turn out to be unfavorable at retirement. We show that investors may mitigate the welfare loss due to annuity risk by conditioning the annuity portfolio on the state of the economy and by hedging certain risks before retirement.

First, we show that it is optimal for individuals to incorporate information on the term structure and risk premia at retirement in the annuity choice. The welfare costs of ignoring this information range from 7% to 9%, depending on the investor's risk preferences. However, the optimal conditional annuitization strategy may depend in a complex way on the underlying state variables. We show that it is possible to design a simple linear rule which allocates wealth to nominal, inflation-linked, and variable annuities contingent on the state of the economy. In fact, 75%-95% of the gains due to incorporating conditioning information can be obtained by following this simple rule. In addition, restricting the annuity menu to only nominal or only inflation-linked annuities increases welfare costs even further. This implies that equity exposure during retirement, as well as the ability to insure inflation risk, are welfare enhancing. These conclusions may have serious implications for both DC and DB pension plans. Concerning DC pension plans, investors tend to restrict attention to nominal annuities, if they annuitize at all. On the other hand, DB pension plans usually offer participants either nominal or inflation-linked annuities. Therefore, the annuity portfolio is not diversified and participants cannot take advantage of the (possibly) investment opportunities offered by equity markets.

Whether or not investors adopt optimal annuitization strategies at retirement, investors face annuity risk before retirement that can be hedged by trading in equity and bond markets. We consider the case where the investor's menu contains stocks, nominal and inflation-linked bonds, and a cash account. We determine the optimal hedging strategies for annuity risk induced by the optimal conditional annuitization strategy, the optimal unconditional annuity strategy, and investing all wealth in inflation-linked or nominal annuities. The optimal hedging strategy entails significant positions in the various securities already in early stages of the investor's life cycle, unless the investor allocates all wealth to inflation-linked annuities. This result is confirmed by a welfare analysis in which we determine the

welfare costs induced by not hedging annuity risk. These welfare costs range from 1% to over 10%, depending on the risk attitude of the investor, unless the investor allocates all wealth to inflation-linked annuities at retirement. In other words, hedging inflation risk and time variation in risk premia before retirement can be significantly welfare enhancing. Hedging annuity risk induced by time variation in real interest rates turns out to be only of minor importance.

Future research can extend this paper in various directions. First, the investor considered annuitizes all wealth at age 65. Neuberger (2003) and Milevsky and Young (2003) have shown that it may be optimal to gradually transfer wealth accumulated to annuities. This relates directly to the literature in which investors endogenously select their retirement age, see for instance Farhi and Panageas (2005) and the references therein. Incorporating this additional flexibility may provide a more complete answer to the dynamic life-cycle investment and consumption problem. Related to this is our assumption that labor is supplied inelastically, implying that the investor does not decide endogenously on the leisure/working decision as in for instance Bodie, Merton, and Samuelson (1992). Second, we have restricted our analysis to immediate individual annuities, ignore bequests, and uncertainty concerning the investor's health during retirement. The annuity menu may be extended by joint annuities for married couples, deferred annuities, or annuities which embed particular derivative structures like escalating annuities. Finally, we have assumed that annuities are priced fairly, while realistic annuity markets usually include charge substantial load factors. Lopes (2005) shows that load factors and, in addition, minimum size restrictions, may have a substantial impact on the annuitization decision.

2.A Pricing of nominal and inflation-linked bonds

We derive the nominal prices of both nominal and inflation-linked bonds in the financial market described in Section 2.2, following the results on affine term structure models in, for instance, Duffie and Kan (1996) and Sangvinatsos and Wachter (2005).

To that extent, we assume that both nominal and inflation-linked bond prices are smooth functions of time and the term structure factors X, which satisfy

$$\mathrm{d}X_t = -K_X \mathrm{d}t + \Sigma_X \mathrm{d}Z_t. \tag{2.33}$$

Denote the price of a nominal bond at time t that matures at time T by $P(X_t, t, T)$. Since nominal bonds are traded assets, we must have that $\phi_t P(X_t, t, T)$ is a martingale, where ϕ is given in (2.7). This implies[20]

$$-P_X K_X X + P_t + \frac{1}{2} tr\left(\Sigma_X' P_{XX} \Sigma_X\right) - RP - P_X' \Sigma_X \Lambda = 0, \tag{2.34}$$

where the subscripts of P denote partial derivatives with respect to the different arguments. Using the results of Duffie and Kan (1996), we obtain nominal bond prices that are exponentially affine in the state variables, like in (2.13). Substituting this expression in (2.34) and matching the coefficients on the constant and the state variables X, we obtain the following set of ordinary differential equations

$$\begin{aligned} \dot{A}(\tau) &= -B(\tau)' \Sigma_X \Lambda_0 + \frac{1}{2} B(\tau)' \Sigma_X \Sigma_X' B(\tau) - \delta_R, & (2.35)\\ \dot{B}(\tau) &= -\left(K_X' + \tilde{\Lambda}_1' \Sigma_X'\right) B(\tau) - \left(\iota_2 - \sigma_\Pi' \tilde{\Lambda}_1\right), & (2.36) \end{aligned}$$

where ι_2 denotes a two dimensional vector of ones and $\tilde{\Lambda}_1$ the first two columns of Λ_1. The boundary conditions to the differential equations are given by $A(0) = 0,\ B(0) = 0$.

The price of inflation-linked bonds can be derived along the same lines. The nominal price of a real bond is denoted by the product $\Pi_t P^R(X, t, T)$. The martingale property of $\phi_t \Pi_t P^R(X_t, t, T)$ implies

$$-P_X^R K_X X + P_t^R + \frac{1}{2} tr\left(\Sigma_X' P_{XX}^R \Sigma_X\right) - (R - \pi + \sigma_\Pi' \Lambda) P^R + P_X^{R\prime} \Sigma_X \left(\sigma_\Pi - \Lambda\right) = 0, \tag{2.37}$$

Since prices of inflation-linked bonds are affine in the state variables,[21] (2.37) reduces to

$$-B^R(\tau)' K_X X - \dot{A}^R(\tau) - \dot{B}^R(\tau)' X + \frac{1}{2} B^R(\tau)' \Sigma_X \Sigma_X' B^R(\tau) - r + B^R(\tau)' \Sigma_X \left(\sigma_\Pi - \Lambda\right) = 0. \tag{2.38}$$

We again match the coefficients on the constant and the state variables X, which leads to the following set of ordinary differential equations

$$\begin{aligned} \dot{A^R}(\tau) &= -B^R(\tau)' \Sigma_X \left(\Lambda_0 - \sigma_\Pi\right) + \frac{1}{2} B^R(\tau)' \Sigma_X \Sigma_X' B^R(\tau) - \delta_r, & (2.39)\\ \dot{B^R}(\tau) &= -\left(K_X' + \tilde{\Lambda}_1' \Sigma_X'\right) B^R(\tau) - e_1, & (2.40) \end{aligned}$$

where e_i denotes the i-th unit vector. The boundary conditions to the differential equations are given by $A^R(0) = 0,\ B^R(0) = 0$.

2.B Details estimation procedure

Our estimation procedure in closely related to Sangvinatsos and Wachter (2005). The main difference is that we allow all yields to be measured with error, following Brennan and Xia (2002) and Campbell and Viceira (2001b), rather than assuming that some yields are measured without error. We assume that the

[20] For notational convenience, we omit the time subscripts.

[21] This is a consequence of the fact that instantaneous expected inflation is affine in the state variables.

measurement errors are independent, both sequentially and cross-sectionally. The continuous time equations underlying the financial market in Section 2.2 can be written as

$$\begin{aligned}
\mathrm{d}\begin{bmatrix} X_t \\ \log \Pi_t \\ \log S_t \\ D_t \end{bmatrix} &= \left(\begin{bmatrix} 0_{2\times 1} \\ \delta_\pi - \frac{1}{2}\sigma_\Pi' \sigma_\Pi \\ \delta_R + \mu_0 - \frac{1}{2}\sigma_S' \sigma_S \\ \mu_D \kappa_D \end{bmatrix} + \begin{bmatrix} -K_X & 0_{2\times 2} & 0 \\ e_2' & 0_{1\times 2} & 0 \\ \left(\iota_2' - \sigma_\Pi' \tilde{\Lambda}_1 + \mu_{1(1:2)}'\right) & 0_{1\times 2} & \mu_{1(3)} \\ 0_{1\times 2} & 0_{1\times 2} & -\kappa_D \end{bmatrix} \begin{bmatrix} X_t \\ \log \Pi_t \\ \log S_t \\ D_t \end{bmatrix} \right) \mathrm{d}t \\
&\quad + \begin{bmatrix} \Sigma_X \\ \sigma_\Pi' \\ \sigma_S' \\ \sigma_D' \end{bmatrix} \mathrm{d}Z_t \\
&= (\Theta_0 + \Theta_1 K_t)\,\mathrm{d}t + \Sigma_K \mathrm{d}Z_t, \qquad (2.41)
\end{aligned}$$

with $K_t = (X_t', \log \Pi_t, \log S_t, D_t)'$ and $K_X \in \mathbb{R}^{2\times 2}$ is a diagonal matrix with elements κ_1 and κ_2. As K_t follows a standard multivariate multivariate Ornstein-Uhlenbeck process, we may write the exact h-period discretization (see for instance Sangvinatsos and Wachter (2005))

$$K_{t+h} = \mu^{(h)} + \Gamma^{(h)} K_t + \varepsilon_{t+h}, \qquad (2.42)$$

where $\varepsilon_{t+h} \overset{i.i.d.}{\sim} N\left(0_{5\times 1}, \Sigma^{(h)}\right)$ for appropriate $\mu^{(h)}$, $\Gamma^{(h)}$, and $\Sigma^{(h)}$ which we derive below. To derive the discrete time parameters, we consider the eigenvalue decomposition $\Theta_1 = UDU^{-1}$. The parameters in the VAR(1) - model relate to the structural parameters via

$$\begin{aligned}
\Gamma^{(h)} &= \exp(\Theta_1 h) = U \exp(Dh)\, U^{-1}, \qquad (2.43) \\
\mu^{(h)} &= \left[\int_t^{t+h} \exp(\Theta_1 [t+h-s])\, ds\right] \Theta_0 \\
&= UFU^{-1}\Theta_0,
\end{aligned}$$

where F is a diagonal matrix with elements $F_{ii} = h\alpha(D_{ii}h)$, with

$$\alpha(x) = \frac{\exp(x) - 1}{x},$$

and $\alpha(0) = 1$. Finally. the derivation of $\Sigma^{(h)}$ is slightly more involved. We have

$$\begin{aligned}
\Sigma^{(h)} &= \int_t^{t+h} \exp(\Theta_1 [t+h-s])\, \Sigma_K \Sigma_K' \exp(\Theta_1 [t+h-s])'\, ds \qquad (2.44) \\
&= UVU',
\end{aligned}$$

where V is a matrix with elements

$$\begin{aligned}
V_{ij} &= \left[\int_t^{t+h} \exp(D[t+h-s])\, U^{-1}\Sigma_K \Sigma_K' \left(U^{-1}\right)' \exp(D[t+h-s])\, ds\right]_{ij} \qquad (2.45) \\
&= \left[U^{-1}\Sigma_K \Sigma_K' \left(U^{-1}\right)'\right]_{ij} \int_t^{t+h} \exp([D_{ii} + D_{jj}][t+h-s])\, ds \\
&= \left[U^{-1}\Sigma_K \Sigma_K' \left(U^{-1}\right)'\right]_{ij} h\alpha([D_{ii} + D_{jj}]h).
\end{aligned}$$

Using data on six yields, stock returns, and inflation, we estimate the model using the Kalman filter. The transition equation is given by (2.42). We assume that all yields are measured with measurement error, in line with Brennan and Xia (2002) and Campbell and Viceira (2001b). The likelihood can subsequently be constructed using the error-prediction decomposition, see for instance Harvey (1989).

2.C Digression on the AIR

In this section, we succinctly summarize the role of the AIR in a simple model. Reducing (2.5), we find

$$\frac{\mathrm{d}S_t}{S_t} = \mu \mathrm{d}t + \sigma \mathrm{d}Z_t, \tag{2.46}$$

with $\sigma = \| \sigma_S \|$ and Z a univariate Brownian motion. For the sake of exposition, we consider in this appendix that the remaining life-time of an individual of age T is exponentially distributed with parameter λ, implying for the survival probabilities

$${}_sp_T = e^{-\lambda s}. \tag{2.47}$$

Therefore, we immediately have, with $s \geq 0$,

$$A^V(h,T) = \frac{1}{\lambda + h}, \tag{2.48}$$

$$I^V(h,T+s,T) = (\lambda+h)\exp\left(\left(\mu - \frac{1}{2}\sigma^2 - h\right)s + \sigma(Z_{T+s} - Z_T)\right), \tag{2.49}$$

see also Charupat and Milevsky (2002). The latter expression reveals that the choice of the AIR affects both the expectation as well as the dispersion of the payments provided by the variable annuity. The expectation and α-quantiles, Q_α, of the payout are given by

$$\mathbb{E}\left(I^V(h,T+s,T)\right) = (\lambda+h)\exp\left((\mu-h)s\right), \tag{2.50}$$

$$Q_\alpha = (\lambda+h)e^{\left(\mu - \frac{1}{2}\sigma^2 - h\right)s} e^{\sqrt{s}\sigma\Phi^{-1}(\alpha)}, \tag{2.51}$$

with Φ denoting the CDF of the standard normal distribution. Obviously, $\mu = h$ is the knife-edge case for which the expected income stream is constant. For $\mu > h$, the initial expected payout is low, but expected to increase during the retirement phase. The early payout is less risky than for $\mu = h$, but increases as the investor ages. On the contrary, for $\mu < h$, the initial payout is high, but expected to decrease. The quantiles for the initial payout will be more wide-spread, but will widen less rapidly as the investor ages. In sum, a low AIR corresponds to low and not too risky initial payout, but future payout is expected to be higher and more risky. A high AIR corresponds to high and relatively risky initial payout, but the future payout is likely to be lower, albeit less risky.

2.D Optimal policies after retirement

We discretize the consumption-savings problem after retirement (2.27) at an annual frequency. We thus consider

$$J(1,Y_T,0,T) = \max_{(C_t)_{t\in(T,\infty)}} \mathbb{E}_T\left(\sum_{t=T+1}^{T_{\max}} {}_{t-T}p_T e^{-\beta(t-T)}\frac{\left(C_t^R\right)^{1-\gamma}}{1-\gamma}\right), \tag{2.52}$$

where we normalize real retirement wealth before annuitization to unity and $T_{\max} = 100$. The maximization is subject to the discretized real budget constraint

$$W_{t+1}^R = (W_t^R - C_t^R)R_{t+1} + Y_{t+1}^R,\ t = T, T+1, \ldots, \text{ and } W_T^R = 0, \tag{2.53}$$

with $Y_t^R = Y_t\Pi_T/\Pi_t$ indicating real annuity income, Y_t nominal annuity income, and $R_{t+1} = \Pi_t/(P(X_t,t,t+1)\Pi_{t+1})$ the real return on a nominal cash account. Further, we impose the liquidity constraint

$$C_t^R \leq W_t^R, \tag{2.54}$$

which implies that the investor cannot capitalize future annuity income to increase today's consumption. Concerning the return on investment, we assume that the investor has access to a cash account, which is

riskless in nominal terms, but risky in real terms.

This problem leads to the following Bellman equation, for $t \geq T$,

$$J(W_t^R, Y_t, 0, t) = \max_{C_t^R} \left\{ \frac{\left(C_t^R\right)^{1-\gamma}}{1-\gamma} + \left(\frac{{}_{t+1}p_T}{{}_tp_T} \right) \mathbb{E}_t \left(e^{-\beta} J(W_{t+1}^R, Y_{t+1}, 0, t+1) \right) \right\}, \tag{2.55}$$

and

$$J(W_{T_{\max}}, Y_{T_{\max}}, 0, T_{\max}) = \frac{\left(C_{T_{\max}}^R\right)^{1-\gamma}}{1-\gamma}. \tag{2.56}$$

Applying the envelope theorem results in the optimal consumption

$$\left(C_t^{R\star}\right)^{-\gamma} = \left(\frac{{}_{t+1}p_T}{{}_tp_T} \right) \mathbb{E}_t \left(e^{-\beta} \left(C_{t+1}^{R\star}\right)^{-\gamma} R_{t+1} \right), \tag{2.57}$$

i.e.,

$$C_t^{R\star} = \left\{ \left(\frac{{}_{t+1}p_T}{{}_tp_T} \right) \mathbb{E}_t \left(e^{-\beta} \left(C_{t+1}^{R\star}\right)^{-\gamma} R_{t+1} \right) \right\}^{-1/\gamma}. \tag{2.58}$$

This result shows that we can determine the optimal consumption policy in a "myopic" fashion. The common procedure is to specify a grid over wealth and subsequently apply numerical dynamic programming to solve for the optimal policy. This results in a strategy $C_{t+1}^{R\star}(W_{t+1}^R)$ which can be used to determine the consumption policy in period t for a given initial wealth level W_t^R. However, solving for the optimal consumption policy then still entails solving for the root of (2.58) as W_{t+1}^R depends on C_t^R. Carroll (2006) provides an alternative approach to circumvent this problem. Carroll (2006) suggests to consider a grid in $a_t = W_t^R - C_t^R$ rather than W_t^R. The budget constraint (2.53) now reads

$$W_{t+1}^R = R_{t+1} a_t + Y_{t+1}^R, \tag{2.59}$$

which no longer depends on C_t^R. As a consequence, once we can approximate the conditional expectation in (2.58), we can solve immediately for the optimal consumption policy

$$C_t^{R\star}(a_t) = \left\{ \left(\frac{{}_{t+1}p_T}{{}_tp_T} \right) \mathbb{E}_t \left(e^{-\beta} \left(C_{t+1}^{R\star}\right)^{-\gamma} R_{t+1} \right) \mid a_t \right\}^{-1/\gamma}. \tag{2.60}$$

Second, once we determine the optimal consumption policy on all m grid points, indicated by $a_1, \ldots, a_m$, we determine an endogenous wealth grid via

$$W_t^R = C_t^{R\star}(a_t) + a_t. \tag{2.61}$$

To approximate the conditional expectations we encounter, we consider expansions of the conditional expectations in a set of basis function of the state variables. This approach has been introduced by Brandt, Goyal, Santa-Clara, and Stroud (2005) for optimal portfolio choice. To review, we first simulate N trajectories and indicate the trajectories by $\omega_1, \ldots, \omega_N$. Second, select a grid of after consumption wealth a, indicated by $a_1, \ldots, a_m$.[22] At time T, the optimal policy is trivial and $C_{T_{\max}}^R = W_{T_{\max}}^R$, $\forall \omega_i$, implying $a_{T_{\max}} = 0$. At times $t = T+1, \ldots, T_{\max}-1$, we estimate the conditional expectation for each a_j using cross-sectional regressions. For wealth levels W_{t+1} in between the (endogenous) grid points, we employ linear interpolation. This leads to the optimal policy for each given a_j. Given all a_j, we determine the endogenous wealth grid W_{jt}^R, $j = 1, \ldots, m$. Along these lines, we proceed backwards. Since the conditional expectation in (2.60) should remain strictly positive, we approximate the conditional expectation not linearly in basis functions, but rather an exponentially affine combination of basis functions. We consider exponentially affine expansions in the real rate, expected inflation, the dividend yield, and log annuity income. The first three state variables are de-meaned and normalized by their (unconditional standard deviation). Higher order expansions hardly

[22] This grid possibly depends on time as well.

change the results at the reported precision.

This numerical procedure results in N trajectories of realized utility. These trajectories are in turn used to estimate the coefficients of the exponentially affine approximation of the value function at retirement. Further details on the numerical procedure are available upon request. Using this approximation, we are able to determine welfare costs of sub-optimal annuitization strategies and the optimal investment and consumption strategies in the period before retirement. More specifically, we consider an approximation of the form[23]

$$[(1-\gamma)J(1,Y_T,0,T)]^{\frac{1}{\gamma-1}} \simeq \exp\left(\xi_0 + \xi_1' Y_T + \frac{1}{2} Y_T' \xi_2 Y_T\right) = K(Y_T). \tag{2.62}$$

Using this approximation, the value function at retirement can be written conveniently as

$$J(W_T^R, Y_T, 0, T) = \frac{1}{1-\gamma}\left(\frac{W_T^R}{K(Y_T)}\right)^{1-\gamma}. \tag{2.63}$$

The utility loss we report (ϕ) is defined as the decrease in certainty equivalent consumption

$$\phi(Y_T) = \left\{\frac{J^{\text{Opt}}(1,Y_T,0,T)}{J^{\text{Sub}}(1,Y_T,0,T)}\right\}^{1/(\gamma-1)} - 1. \tag{2.64}$$

Alternatively, $\phi(Y_T)$ can be interpreted as the fraction of retirement wealth an investor is willing to give up in order to be able to implement the optimal conditional annuitization strategy. Since this measure depends on the state of the economy at retirement, we report its unconditional expectation.[24]

2.E Optimal policies before retirement

In this appendix we derive the optimal investment strategy in the period before retirement using dynamic programming. We initially solve the problem without labor income. The approximate objective function is given by

$$\mathbb{E}_0\left\{\frac{1}{1-\gamma}\left(\frac{W_T^R}{K(Y_T)}\right)^{1-\gamma}\right\}, \tag{2.65}$$

with $W_T^R = W_T \Pi_T^{-1}$, subject to the dynamic budget constraint of real wealth

$$\begin{aligned}\frac{\mathrm{d}W_t^R}{W_t^R} &= \left(\delta_r + e_1' Y_t + \sigma_\Pi'\left(\sigma_\Pi - \Lambda_t\right) + x_t'\Sigma\left(\Lambda_t - \sigma_\Pi\right)\right)\mathrm{d}t + \left(x_t'\Sigma - \sigma_\Pi'\right)\mathrm{d}Z_t &(2.66)\\ &= \mu_{W^R}\mathrm{d}t + \sigma_{W^R}'\mathrm{d}Z_t, &(2.67)\end{aligned}$$

with e_i denoting the i-th unit vector and $Y = (X_1, X_2, D)'$. The diffusion of the state vector is given by

$$\mathrm{d}Y_t = \left(\zeta_0 + \zeta_1 Y_t\right)\mathrm{d}t + \Sigma_Y \mathrm{d}Z_t, \tag{2.68}$$

with

$$\zeta_0 = \begin{pmatrix} 0 \\ 0 \\ \kappa_D \mu_D \end{pmatrix}, \zeta_1 = \begin{pmatrix} -\kappa_1 & 0 & 0 \\ 0 & -\kappa_2 & 0 \\ 0 & 0 & -\kappa_D \end{pmatrix}, \Sigma_Y = \begin{pmatrix} \sigma_1' \\ \sigma_2' \\ \sigma_D' \end{pmatrix}. \tag{2.69}$$

[23] This approximation can be shown to be highly accurate. However, the accuracy deteriorates for high risk aversion levels and annuity portfolios which are concentrated in nominal annuities. Results on the accuracy of the value function are available upon request.

[24] We use a lemma in Campbell, Chan, and Viceira (2003) to determine $\mathbb{E}\left(\phi(Y_T)\right)$.

Using the definition of $K(Y)$ in (2.62), the dynamics of $K_t = K(Y_t)$ is given by

$$\begin{aligned}\frac{\mathrm{d}K_t}{K_t} &= \left((\xi_1+\xi_2 Y_t)'(\zeta_0+\zeta_1 Y_t)+\frac{1}{2}tr\left(\Sigma_Y'\xi_2\Sigma_Y\right)+\frac{1}{2}(\xi_1+\xi_2 Y_t)'\Sigma_Y\Sigma_Y'(\xi_1+\xi_2 Y_t)\right)\mathrm{d}t \\ &\quad +(\xi_1+\xi_2 Y_t)'\Sigma_Y\mathrm{d}Z_t \qquad (2.70)\\ &= \mu_K\mathrm{d}t+\sigma_K'\mathrm{d}Z_t. \qquad (2.71)\end{aligned}$$

We first of all derive the dynamics of wealth relative to K_t, which we denote by W^K. Next, we derive the optimal portfolio and the induced value function before retirement. Importantly, we show in (2.82) that the optimal portfolio policy is affine in the state variables, Y, and we introduce the notation

$$x_t = \alpha_0+\alpha_1 Y_t. \qquad (2.72)$$

The dynamics of scaled real wealth, W^K, is given by

$$\frac{\mathrm{d}W_t^K}{W_t^K} = \left(\mu_{W^K}^C+\mu_{W^K}^Y Y_t+Y_t'\mu_{W^K}^{YY}Y_t\right)\mathrm{d}t+\left(\sigma_{W^K}^C+\sigma_{W^K}^Y Y_t\right)'\mathrm{d}Z_t. \qquad (2.73)$$

with drift coefficients

$$\begin{aligned}\mu_{W^K}^C &= \delta_r+(\Sigma'\alpha_0-\sigma_\Pi)'(\Lambda_0-\sigma_\Pi)-\xi_1'\zeta_0-\frac{1}{2}tr\left(\Sigma_Y\xi_2\Sigma_Y'\right)+\frac{1}{2}\xi_1'\Sigma_Y\Sigma_Y'\xi_1- \\ &\quad \xi_1'\Sigma_Y(\Sigma'\alpha_0-\sigma_\Pi), \qquad (2.74)\\ \mu_{W^K}^Y &= e_1'-\sigma_\Pi'\Lambda_1+\alpha_0'\Sigma\Lambda_1+(\Lambda_0-\sigma_\Pi)'\Sigma'\alpha_1-\xi_1'\zeta_1-\zeta_0'\xi_2+\xi_1'\Sigma_Y\Sigma_Y'\xi_2- \\ &\quad \xi_1'\Sigma_Y\Sigma'\alpha_1-(\alpha_0'\Sigma-\sigma_\Pi')\Sigma_Y'\xi_2, \qquad (2.75)\\ \mu_{W^K}^{YY} &= \alpha_1'\Sigma\Lambda_1-\xi_2'\zeta_1+\frac{1}{2}\xi_2'\Sigma_Y\Sigma_Y'\xi_2-\xi_2\Sigma_Y\Sigma'\alpha_1, \qquad (2.76)\end{aligned}$$

and diffusion coefficients

$$\begin{aligned}\sigma_{W^K}^C &= \Sigma'\alpha_0-\sigma_\Pi-\Sigma_Y'\xi_1, \qquad (2.77)\\ \sigma_{W^K}^Y &= \Sigma'\alpha_1-\Sigma_Y'\xi_2. \qquad (2.78)\end{aligned}$$

The value function is conjectured to be of the form

$$J(W_t^R,Y_t,t) = \frac{1}{1-\gamma}\left(\frac{W_t^R}{K(Y_t)}\right)^{1-\gamma}\exp\left(\Gamma_0(\tau)+\Gamma_1(\tau)'Y_t+\frac{1}{2}Y_t'\Gamma_2(\tau)Y_t\right), \qquad (2.79)$$

with $\tau = T-t$ the remaining time up to retirement. The optimal investment policy is subsequently derived via the Hamilton-Jacobi-Bellman (HJB) equation

$$\sup_x\begin{pmatrix} J_{W^K}W^K\left(\mu_{W^K}^C+\mu_{W^K}^Y Y_t+Y_t'\mu_{W^K}^{YY}Y_t\right)+J_{W^KY}'\Sigma_Y\left(\sigma_{W^K}^C+\sigma_{W^K}^Y Y_t\right)+ \\ \frac{1}{2}J_{W^KW^K}W^{K2}\left(\sigma_{W^K}^C+\sigma_{W^K}^Y Y_t\right)'\left(\sigma_{W^K}^C+\sigma_{W^K}^Y Y_t\right)+J_Y'(\zeta_0+\zeta_1 Y)+ \\ \frac{1}{2}tr\left(\Sigma_Y'J_{YY}\Sigma_Y\right)+J_t\end{pmatrix}=0, \qquad (2.80)$$

subject to the boundary condition

$$J(W_T^R,Y_T,T) = \frac{1}{1-\gamma}\left(\frac{W_T^R}{K(Y_T)}\right)^{1-\gamma}. \qquad (2.81)$$

Using the first order conditions of (2.80) and the value function as given in (2.79), the optimal investment strategy is given by

$$\begin{aligned}x_t^\star &= \frac{1}{\gamma}(\Sigma\Sigma')^{-1}\Sigma\Lambda_t+\left(1-\frac{1}{\gamma}\right)(\Sigma\Sigma')^{-1}\Sigma(\sigma_\Pi+\Sigma_Y'(\xi_1+\xi_2 Y_t))+ \\ &\quad \frac{1}{\gamma}(\Sigma\Sigma')^{-1}\Sigma\Sigma_Y'\left(\Gamma_1(\tau)+\frac{1}{2}(\Gamma_2(\tau)+\Gamma_2(\tau)')Y_t\right), \qquad (2.82)\end{aligned}$$

so that the coefficients in (2.72) are given by

$$\begin{aligned}\alpha_0 &= \frac{1}{\gamma}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Lambda_0 + \left(1-\frac{1}{\gamma}\right)\left(\Sigma\Sigma'\right)^{-1}\Sigma\left(\sigma_\Pi + \Sigma_Y'\xi_1\right) + \\ &\quad \frac{1}{\gamma}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Sigma_Y'\Gamma_1(\tau), \qquad (2.83)\\ \alpha_1 &= \frac{1}{\gamma}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Lambda_1 + \left(1-\frac{1}{\gamma}\right)\left(\Sigma\Sigma'\right)^{-1}\Sigma\Sigma_Y'\xi_2 + \\ &\quad \frac{1}{2}\frac{1}{\gamma}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Sigma_Y'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right). \qquad (2.84)\end{aligned}$$

Substitution of the optimal policy into the HJB-equation (2.80) implies for the coefficients of the value function

$$\begin{aligned}\dot{\Gamma_0}(\tau) &= (1-\gamma)\mu_{W^K}^C - \frac{1}{2}\gamma(1-\gamma)\sigma_{W^K}^{C\prime}\sigma_{W^K}^C + \Gamma_1(\tau)'\zeta_0 + \frac{1}{2}\Gamma_1(\tau)'\Sigma_Y\Sigma_Y'\Gamma_1(\tau) + \\ &\quad \frac{1}{4}tr\left(\Sigma_Y'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right)\Sigma_Y\right) + (1-\gamma)\Gamma_1(\tau)'\Sigma_Y\sigma_{W^K}^C, \qquad (2.85)\\ \dot{\Gamma_1}(\tau)' &= (1-\gamma)\mu_{W^K}^Y - \gamma(1-\gamma)\sigma_{W^K}^{C\prime}\sigma_{W^K}^Y + \Gamma_1(\tau)'\zeta_1 + \frac{1}{2}\zeta_0'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right) + \\ &\quad \frac{1}{2}\Gamma_1(\tau)'\Sigma_Y\Sigma_Y'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right) + (1-\gamma)\Gamma_1(\tau)'\Sigma_Y\sigma_{W^K}^Y + \\ &\quad \frac{1}{2}(1-\gamma)\sigma_{W^K}^{C\prime}\Sigma_Y'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right) \qquad (2.86)\\ \dot{\Gamma_2}(\tau) &= 2(1-\gamma)\mu_{W^K}^{YY} - \gamma(1-\gamma)\sigma_{W^K}^{Y\prime}\sigma_{W^K}^Y + \left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right)\zeta_1 + \\ &\quad \frac{1}{4}\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right)\Sigma_Y\Sigma_Y'\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right) + \\ &\quad (1-\gamma)\left(\Gamma_2(\tau)+\Gamma_2(\tau)'\right)\Sigma_Y\sigma_{W^K}^Y, \qquad (2.87)\end{aligned}$$

subject to the boundary conditions $\Gamma_0(0)=0$, $\Gamma_1(0)=0$, $\Gamma_2(0)=0$. Note that the value function corresponding to sub-optimal strategies satisfies the same ODEs, although the expressions for α_0 and α_1 should be modified.

In order to solve the investment problem with labor income, Bodie, Merton, and Samuelson (1992) and Munk and Sørensen (2005) have shown that it is possible to recast this problem to one without labor income and initial real wealth $W_t^R + H_t$. On the one hand, the optimal exposures to the risk factors of total real wealth at time t are given by $(W_t^R + H_t)(x_t^{\star\prime}\Sigma - \sigma_\Pi')$. On the other hand, for any investment strategy, $\bar{x}_t$, we can derive another expression for the diffusion coefficient of real total wealth using (2.20) together with the dynamic of realized inflation in (2.4) to determine the dynamics of real financial wealth and (2.31) for the dynamics of the real present value of the savings stream, namely $W_t^R\left(\bar{x}_t'\Sigma - \sigma_\Pi\right) + H_{Y_t}'\Sigma_Y$. By equating both diffusion coefficients and solving for the optimal portfolio $\bar{x}_t$, we immediately obtain the expression provided in (2.32).[25]

Next, we solve for the optimal fixed consumption rate, θ. The value function at time $t = t_0$ induced by the optimal investment strategy is given by

$$\begin{aligned}J(W_{t_0}, Y_{t_0}, L_{t_0}, t_0) &= \max_\theta \int_{t_0}^T e^{-\beta s}\frac{(\theta L_s)^{1-\gamma}}{1-\gamma}\mathrm{d}s + \frac{e^{-\beta T}}{1-\gamma}\left(\frac{W_{t_0}^R + H(L_{t_0}, Y_{t_0}, \theta, t_0, T)}{K(Y_{t_0})}\right)^{1-\gamma} \times \\ &\quad \exp\left(\Gamma_0(\tau) + \Gamma_1(\tau)'Y_{t_0} + \frac{1}{2}Y_{t_0}'\Gamma_2(\tau)Y_{t_0}\right). \qquad (2.88)\end{aligned}$$

where $\tau = T - t_0$. Using the definition of H_t in (2.31), it follows immediately that the optimal consumption

[25] In the main text, the optimal portfolio in the presence of labor income is indicated by $x_t^\star$, with slight abuse of notation.

rate is given by

$$\theta^\star = \frac{\left(\frac{\rho_1}{\rho_2}\right)^{\frac{1}{\gamma}}}{1+\left(\frac{\rho_1}{\rho_2}\right)^{\frac{1}{\gamma}}}, \tag{2.89}$$

with, assuming $W^R_{t_0} = 0$,

$$\rho_1 = \int_{t_0}^{T} e^{-\beta s} \left(L_s\right)^{1-\gamma} \mathrm{d}s, \tag{2.90}$$

$$\rho_2 = e^{-\beta T} \left(\frac{H(L_{t_0}, Y_{t_0}, 0, t_0, T)}{K(Y_{t_0})}\right)^{1-\gamma} \exp\left(\Gamma_0(\tau) + \Gamma_1(\tau)' Y_{t_0} + \frac{1}{2} Y'_{t_0} \Gamma_2(\tau) Y_{t_0}\right), \tag{2.91}$$

and $\tau = T - t_0$. We will assume throughout that initial wealth, i.e. $W^R_{t_0}$, equals zero.

Finally, using the expressions for the value function in (2.88), we can determine the fraction of retirement wealth an investor is willing to give up to hedge annuity risk is given by, with $\tau = T - t_0$,

$$\phi(Y_{t_0}) = \tag{2.92}$$
$$\exp\left(\Gamma_0^{\mathrm{Opt}}(\tau) - \Gamma_0^{\mathrm{Sub}}(\tau) + \left(\Gamma_1^{\mathrm{Opt}}(\tau) - \Gamma_1^{\mathrm{Sub}}(\tau)\right)' Y_{t_0} + \frac{1}{2} Y'_{t_0} \left(\Gamma_2^{\mathrm{Opt}}(\tau) - \Gamma_2^{\mathrm{Sub}}(\tau)\right) Y_{t_0}\right)^{1/(\gamma-1)} - 1,$$

where the coefficients with superscripts 'Opt' originate from the value function corresponding to the optimal strategy which anticipates the investor's desire to annuitize wealth at retirement. Likewise, the superscripts 'Sub' correspond to the value function generated by a strategy which perceives ξ_1 and ξ_2 to be equal to zero, see (2.32). The welfare costs are calculated assuming that the initial state vector, Y_{t_0}, equals its unconditional expectation.

2.F Tables and figures

Parameter	Estimate	Parameter	Estimate
Means nominal short rate and expected inflation: $\mathbb{E}(R_t) = \delta_R$, $\mathbb{E}(\pi_t) = \delta_\pi$			
δ_R	4.69%	δ_π	3.02%
Process real interest rate and expected inflation: $\mathrm{d}X_{it} = -\kappa_i X_{it}\mathrm{d}t + \sigma_i'\mathrm{d}Z_t$			
κ_1	1.07	σ_1	1.98%
κ_2	0.08	σ_{12}	-0.13%
		σ_2	1.05%
Realized inflation process: $\mathrm{d}\Pi_t/\Pi_t = \pi_t\mathrm{d}t + \sigma_\Pi'\mathrm{d}Z_t$			
$\sigma_{\Pi(1)}$	0.10%	$\sigma_{\Pi(3)}$	1.08%
$\sigma_{\Pi(2)}$	0.18%		
Stock return process: $\mathrm{d}S_t/S_t = (R_t + \mu_0 + \mu_1'Y_t)\mathrm{d}t + \sigma_S'\mathrm{d}Z_t$			
μ_0	0.44	$\sigma_{S(1)}$	-1.50%
$\mu_{1(1)}$	-0.56	$\sigma_{S(2)}$	-2.49%
$\mu_{1(2)}$	-0.80	$\sigma_{S(3)}$	-1.53%
$\mu_{1(3)}$	0.11	$\sigma_{S(4)}$	14.48%
Prices of real rate and inflation risk: $\Lambda_t = \Lambda_0 + \Lambda_1 Y_t$			
$\Lambda_{0(1)}$	-0.35	$\Lambda_{1(1,1)}$	-26.01
$\Lambda_{0(2)}$	-0.13	$\Lambda_{1(2,2)}$	-7.07
Dividend yield process: $\mathrm{d}D_t = \kappa_D(\mu_D - D_t)\mathrm{d}t + \sigma_D'\mathrm{d}Z_t$			
κ_D	0.05	$\sigma_{D(3)}$	1.48%
μ_D	-3.50	$\sigma_{D(4)}$	-14.40%
$\sigma_{D(1)}$	1.73%	$\sigma_{D(5)}$	3.73%
$\sigma_{D(2)}$	2.44%		

Table 2.1: Estimation results for the financial market model
Parameter estimates of the financial market model. The model is estimated using monthly data on six bond yields, inflation, and stock returns over the period from January 1952 up to May 2002. Details on the estimation procedure are provided in Appendix 2.B.

$\gamma = 2$	Constant	Real rate	Exp. inflation	Div. yield
Nominal annuity	12%	-1%	12%	-10%
Inflation-linked annuity	23%	2%	2%	-26%
Variable annuity	65%	-2%	-14%	36%
$\gamma = 5$	Constant	Real rate	Exp. inflation	Div. yield
Nominal annuity	8%	0%	8%	-1%
Inflation-linked annuity	50%	2%	3%	-27%
Variable annuity	42%	-2%	-11%	28%
$\gamma = 10$	Constant	Real rate	Exp. inflation	Div. yield
Nominal annuity	6%	0%	5%	2%
Inflation-linked annuity	70%	1%	2%	-20%
Variable annuity	25%	-1%	-7%	18%

Table 2.2: Optimal retirement choice
Optimal annuity choice *conditional* on the economic conditions at retirement. The optimal allocation to nominal, inflation-linked, and variable annuities with an AIR of $h = 4\%$ is presented. We present the coefficients of a regression of the optimal allocation to the three annuity products on the three state variables. The time preference parameter equals $\beta = 0.04$ and the coefficient of relative risk aversion equals $\gamma =$ 2, 5, or 10. The main text provides further details.

	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
Nominal annuity	0%	0%	0%
Inflation-linked annuity	30%	60%	75%
Variable annuity	70%	40%	25%

Table 2.3: Optimal unconditional retirement choice
Optimal *unconditional* annuity choice at retirement. The optimal allocation to nominal, inflation-linked, and variable annuities with an AIR of $h = 4\%$ is presented. The time preference parameter equals $\beta = 0.04$ and the coefficient of relative risk aversion equals $\gamma = 2$, 5, or 10. The main text provides further details.

	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
Optimal unconditional	-8.91%	-8.64%	-6.88%
Inflation-linked	-18.85%	-13.89%	-9.91%
Nominal	-22.38%	-27.51%	-54.84%
Linear rule	-2.10%	-0.47%	-0.29%

Table 2.4: Welfare costs of sub-optimal annuitization strategies
Welfare costs of *sub-optimal* annuitization strategies relative to the optimal conditional annuitization strategy. Welfare costs are determined as the decrease in certainty equivalent consumption during retirement. As this welfare metric depends on the state of the economy, we report unconditional expectations. The sub-optimal annuitization strategies are either not using conditioning information ('Optimal unconditional'), investing all wealth in either inflation-linked ('Inflation-linked') or nominal ('Nominal') annuities, or the linear rule based on the first order approximation of the optimal conditional annuitization strategy. The time preference parameter equals $\beta = 0.04$ and the coefficient of relative risk aversion equals $\gamma = 2$, 5, or 10. The main text provides further details.

	$\gamma = 2$	$\gamma = 5$	$\gamma = 10$
Optimal conditional	-2.03%	-9.20%	-11.98%
Optimal unconditional	-2.43%	-9.12%	-13.35%
Inflation-linked	-0.02%	-0.13%	-0.35%
Nominal	-0.10%	-0.59%	-1.79%

Table 2.5: Welfare costs of not hedging annuity risk before retirement
Welfare costs of not hedging annuity risk before retirement. Welfare costs are determined as the amount of wealth an investor is willing to give up in order to follow the optimal investment strategy. The reported numbers are based on the condition that the initial vector of state variables, Y_{t_0}, equals its unconditional expectation. Apart from the optimal annuitization strategy ('Optimal conditional'), the sub-optimal annuitization strategies either ignore information on the term structure and risk premia ('Optimal unconditional') or invest all wealth in either inflation-linked ('Inflation-linked') or nominal ('Nominal') annuities. The state variables are set to their unconditional expectation. The time preference parameter equals $\beta = 0.04$ and the coefficient of relative risk aversion ranges from $\gamma = 2$, 5, or 10.

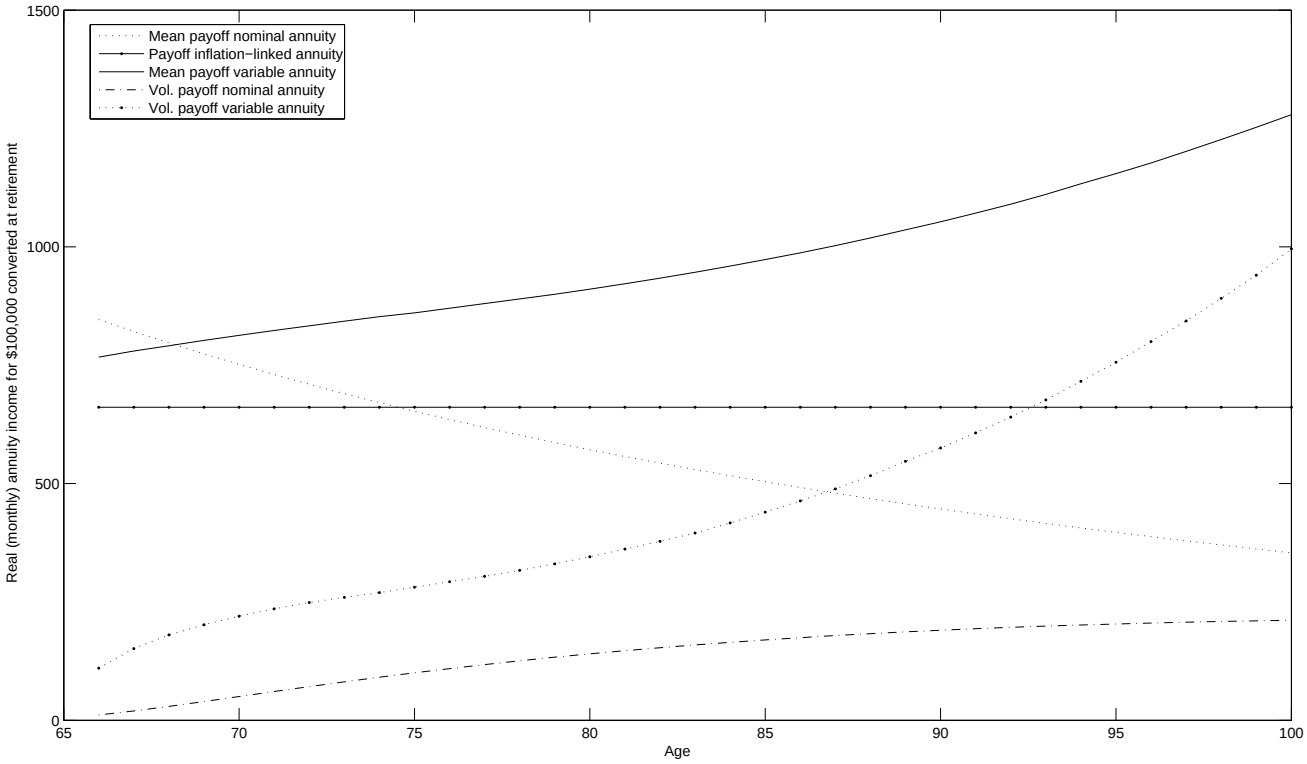


Figure 2.1: Mean and volatility of real (monthly) annuity benefits provided by various annuities
Mean and volatility of real (monthly) annuity benefits provided by nominal, inflation-linked, and variable annuities ($h = 4\%$) for \$100,000 converted at retirement. The real income stream provided by an inflation-linked annuity is by definition constant and therefore only its level is reported. The horizontal axis depicts the investor's age and the vertical axis indicates the level of the real annuity payout. The state variables equal their unconditional expectation.

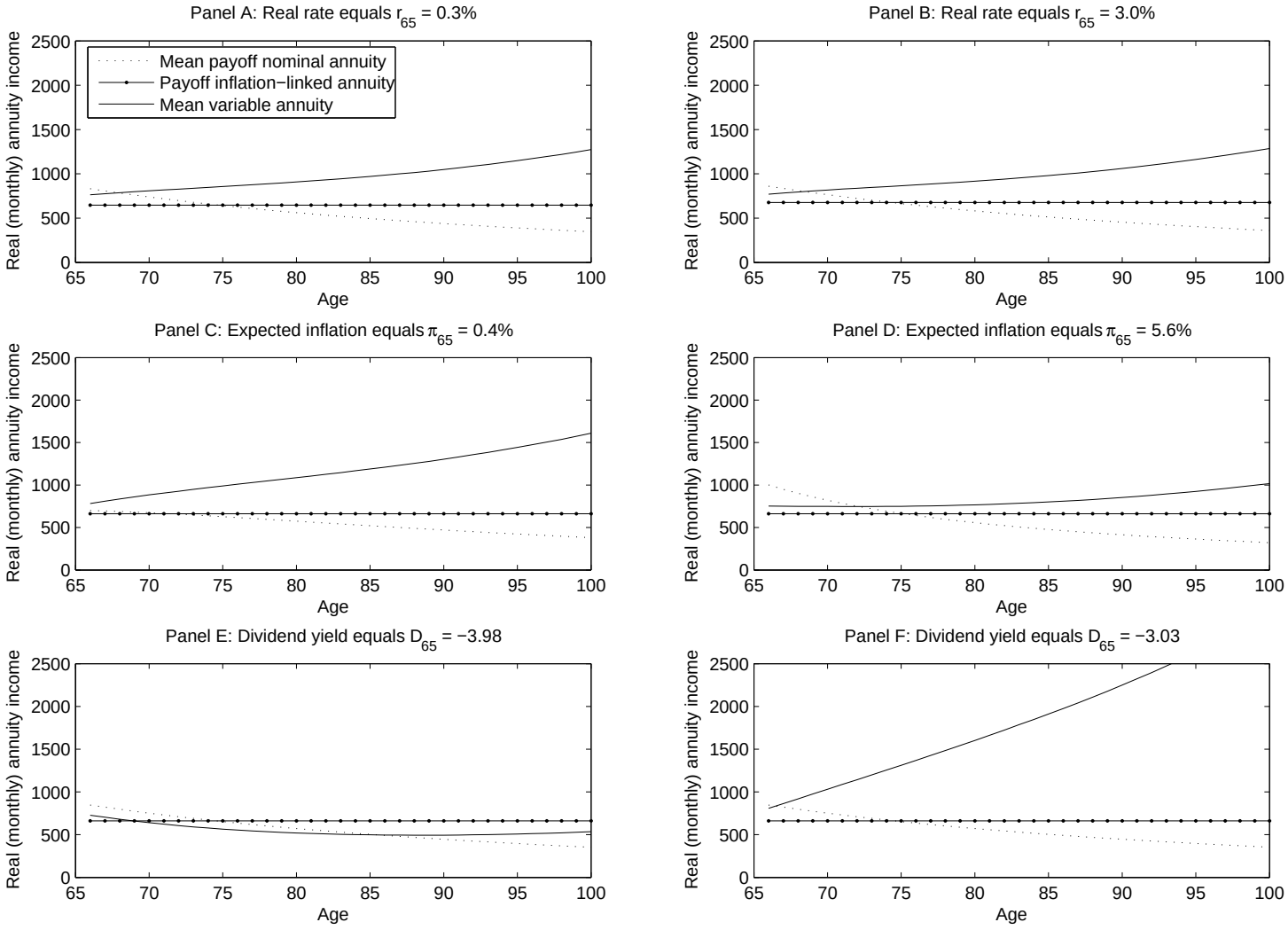


Figure 2.2: Mean real (monthly) annuity benefits provided by nominal, inflation-linked, and variable annuities for different economic conditions

Mean real (monthly) annuity benefits provided by nominal, inflation-linked, and variable annuities for various economic conditions at retirement if the investor converts \$100,000 into annuities. The top panels portray the characteristics of the income stream when the real rate is either one (unconditional) standard deviation below (Panel A) or above (Panel B) its unconditional expectation. The middle panels display the results when expected inflation is either one (unconditional) standard deviation below (Panel C) or above (Panel D) its unconditional expectation. Likewise, the bottom panels present the characteristics of the income stream for an initial level of the dividend yield which is one (unconditional) standard deviation below (Panel E) or above (Panel F) its unconditional expectation. The horizontal axis depicts the investor's age and the vertical axis indicates the level of the real annuity payout.

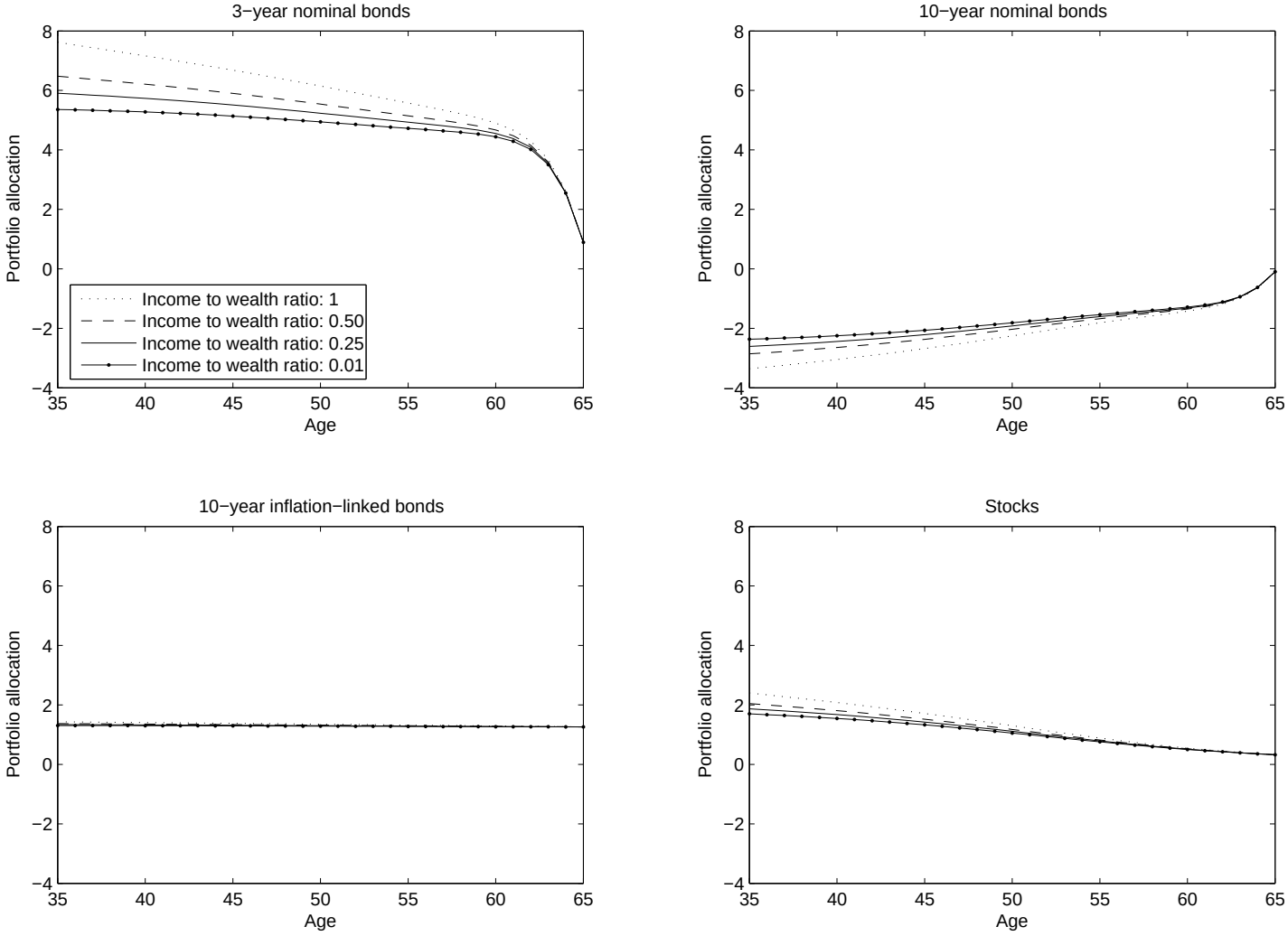


Figure 2.3: Optimal portfolio choice before retirement without annuity risk
Optimal portfolio choice before retirement without annuity risk using 3-year and 10-year nominal bonds, 10-year inflation-linked bonds, and stocks. The remainder is invested in a nominal cash account. The different lines in each of the graphs correspond to income to wealth ratios (L_t/W_t^R) of 1, 0.5, 0.25, and 0.01. The investor's coefficient of relative risk aversion equals $\gamma = 5$. In all graphs, the horizontal axis depicts the investor's age and the vertical axis indicates the optimal allocation to a particular asset. The main text provides further details.

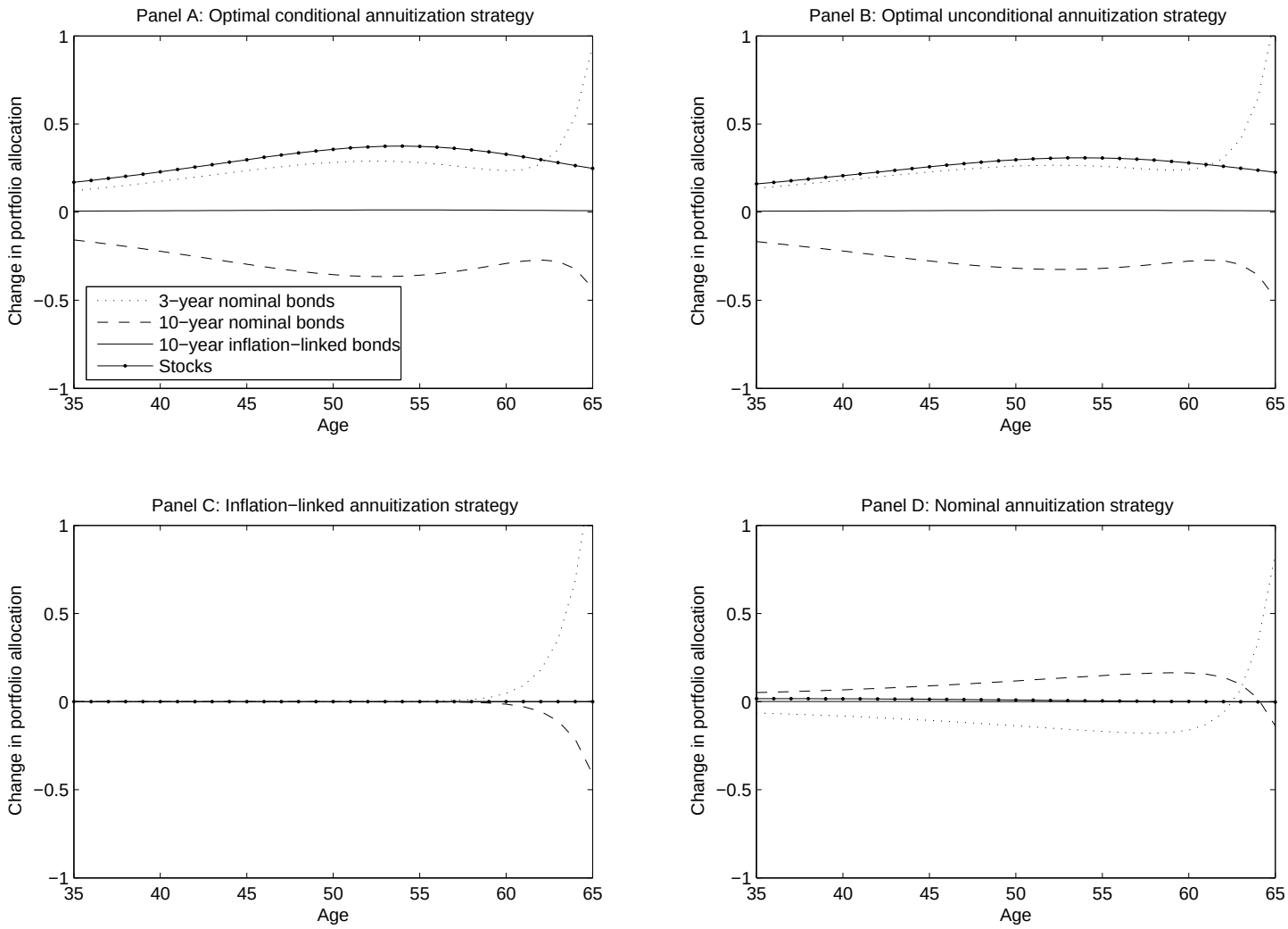


Figure 2.4: Optimal hedging strategy before retirement corresponding to four annuitization strategies

Optimal hedging strategy before retirement to hedge the annuity risk caused by the optimal conditional annuitization strategy (Panel A), the optimal unconditional annuitization strategy (Panel B), inflation-linked annuitization (Panel C), and finally nominal annuitization (Panel D). The optimal hedging strategy is defined the difference between the optimal investment strategy which does and does not account for annuity risk. The asset menu contains 3-year and 10-year nominal bonds, 10-year inflation-linked bonds, and stocks. The remainder is invested in a nominal cash account. The income-to-wealth ratio is set to 0.5 and the investor's coefficient of relative risk aversion equals $\gamma = 5$. In all graphs, the horizontal axis depicts the investor's age and the vertical axis indicates the optimal allocation to a particular asset.

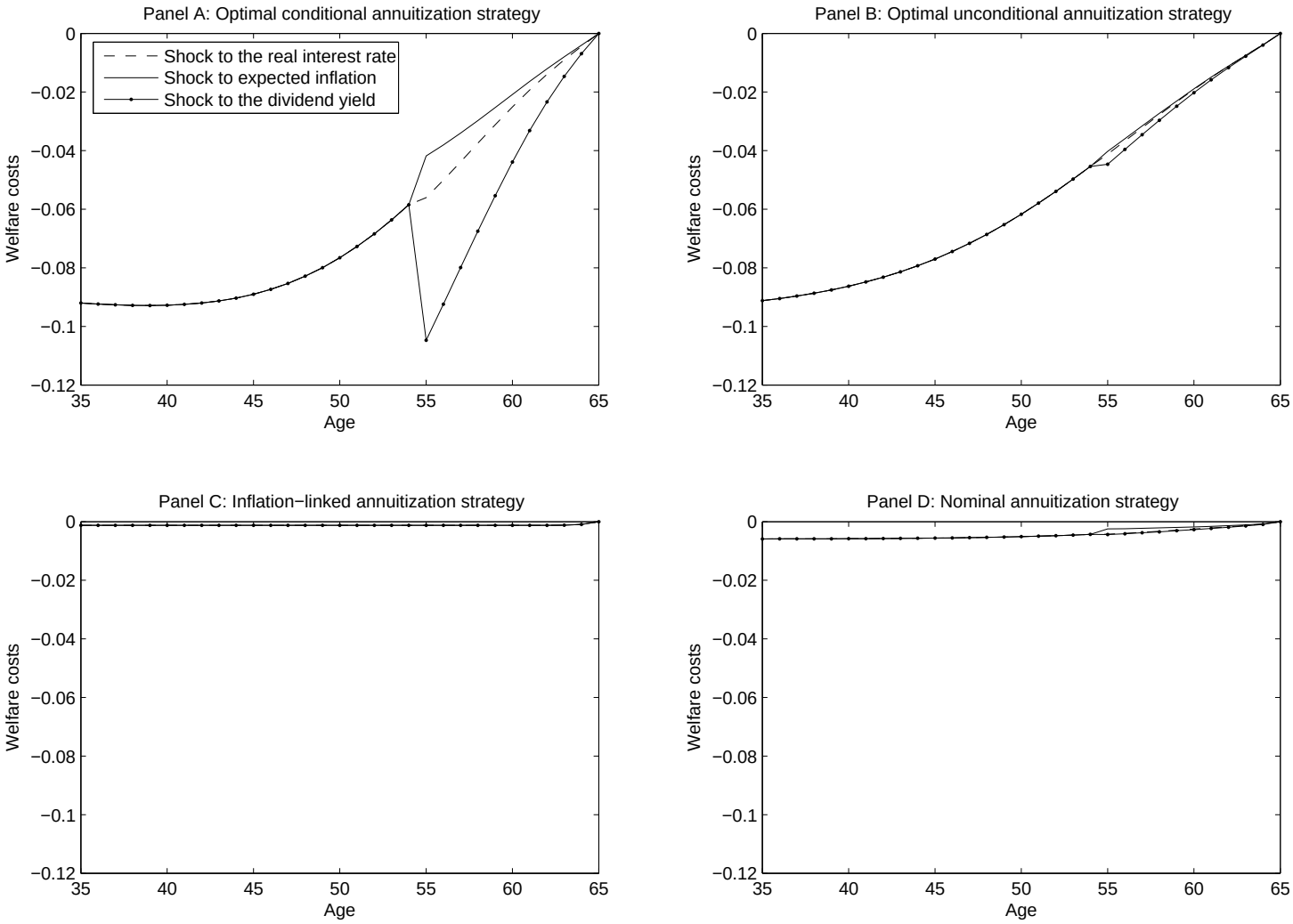


Figure 2.5: Welfare costs of not hedging annuity risk before retirement for different state variables

Welfare costs of not hedging annuity risk before retirement. Welfare costs are determined as the amount of wealth an investor is willing to give up in order to follow the optimal investment strategy. The reported numbers are based on the condition that the initial vector of state variables, Y_{t_0}, equals its unconditional expectation. Apart from the optimal annuitization strategy ('Optimal conditional'), the sub-optimal annuitization strategies either ignore information on the term structure and risk premia ('Optimal unconditional') or invest all wealth in either inflation-linked ('Inflation-linked') or nominal ('Nominal') annuities. The state variables are set to their unconditional expectation up to age 55, perturbed by a one (unconditional) standard deviation, positive shock at age 55, and their expected value, conditional upon the shock, afterwards. The time preference parameter equals $\beta = 0.04$ and the coefficient of relative risk aversion ranges from $\gamma = 2$, 5, or 10. The main text provides further details.

Chapter 3

Optimal Decentralized Investment Management

Abstract

We study an institutional investment problem in which a centralized decision maker, the Chief Investment Officer (CIO), for example, employs multiple asset managers to implement and execute investment strategies in separate asset classes. The CIO allocates capital to the managers who, in turn, allocate these funds to the assets in their asset class. This two-step investment process causes several misalignments of objectives between the CIO and his managers and can lead to large utility costs on the part of the CIO. We focus on (i) loss of diversification, (ii) unobservable appetites for risk of the managers, and (iii) different investment horizons. We derive an optimal unconditional linear performance benchmark and show that this benchmark can be used to better align incentives within the firm. We find that the CIOs uncertainty about the managers risk appetites increases both the costs of decentralized investment management and the value of an optimally designed benchmark.

3.1 Introduction

The investment management divisions of banks, mutual funds, and pension funds are predominantly structured around asset classes such as equities, fixed income, and alternative investments. To achieve superior returns, either through asset selection or market timing, gathering information about specific assets and capitalizing on the acquired informational advantage requires a high level of specialization. This induces the centralized decision maker of the firm, the Chief Investment Officer (CIO), for example, to pick asset managers who are specialized in a single asset class and to delegate portfolio decisions to these specialists. As a consequence, asset allocation decisions are made in at least two stages. In the first stage, the CIO allocates capital to the different asset classes, each managed by a different asset manager. In the second stage, each manager decides how to allocate the funds made available to him, that is, to the assets within his class. This two-stage process can generate several misalignments of incentives that may lead to large utility costs on the part of the

CIO. We show that designing appropriate return benchmarks can substantially reduce these costs.

We focus on the following important, although not exhaustive, list of misalignments of incentives. First, the two-stage process can lead to severe diversification losses. The unconstrained (single-step) solution to the mean-variance optimization problem is likely different from the optimal linear combination of mean-variance efficient portfolios in each asset class, as pointed out by Sharpe (1981) and Elton and Gruber (2004). Second, there may be considerable, but unobservable, differences in appetites for risk between the CIO and each of the asset managers. When the CIO only knows the cross-sectional distribution of risk appetites of investment managers but does not know where in this distribution a given manager falls, delegating portfolio decisions to multiple managers can be very costly. Third, the investment horizons of the asset managers and of the CIO may be different. Since the managers are usually compensated on an annual basis, their investment horizon is generally relatively short. The CIO, in contrast, may have a much longer investment horizon.

In practice, the performance of each asset manager is measured against a benchmark comprised of a large number of assets within his class. In the literature, the main purpose of these benchmarks has been to disentangle the effort and achievements of the asset manager from the investment opportunity set available to him. In this paper we show that an optimally designed unconditional benchmark can also serve to improve the alignment of incentives within the firm and to substantially mitigate the utility costs of decentralized investment management.

Our results provide a different perspective on the use of performance benchmarks. Admati and Pfleiderer (1997) take a realistic benchmark as given and show that when an investment manager uses the conditional return distribution in his investment decisions, restricting him by an unconditional benchmark distorts incentives.[1] In their framework, this distortion can only be prevented by setting the benchmark equal to the minimum-variance portfolio. We show that the negative aspect of unconditional benchmarks can be offset, at least in part, by the role of unconditional benchmarks in aligning other incentives, such as diversification, risk preferences, and investment horizons.

We use a stylized representation of an investment management firm to quantify the costs of the misalignments for both constant and time-varying investment opportunities. We assume that the CIO acts in the best interest of a large group of beneficiaries of the assets under management, whereas the investment managers only wish to maximize their personal compensation. Using two asset classes (bonds and stocks) and three assets per class (government bonds, Baa-rated corporate bonds, and Aaa-rated corporate bonds in the

[1] See also Basak, Shapiro, and Tepla (2006).

fixed income class, and growth stocks, intermediate, and value stocks in the equities class) the utility costs can range from 50 to 300 basis points per year. We therefore argue that decentralization has a first-order effect on the performance of investment management firms.

We demonstrate that when the investment opportunity set is constant and risk attitudes are observable, the CIO can fully align incentives through an unconditional benchmark consisting only of assets in each manager's asset class. In other words, cross-benchmarking is not required. Furthermore, we derive the perhaps counterintuitive result that the risk aversion levels of the asset managers for which the utility costs of the CIO are minimized can substantially differ from the risk aversion of the CIO. We then consider the case of time-varying investment opportunities and show that an unconditional (passive) benchmark can still substantially, though not fully, mitigate the utility costs of decentralized investment management.

Next we generalize our model by relaxing the assumption that the CIO knows the asset managers' risk appetite. Specifically, we derive the optimal benchmark assuming that the CIO only knows the cross-sectional distribution of investment managers' risk aversion levels but does not know where in this distribution a given manager falls. We find that the qualitative results on the benefits of optimal benchmarking derived for a known risk aversion level apply to this more general case. In fact, we find that uncertainty about the managers' risk appetites increases both the costs of decentralized investment management and the value of an optimally designed benchmark.

The negative impact of decentralized investment management on diversification was first noted by Sharpe (1981), who shows that if the CIO has rational expectations about the portfolio choices of the investment managers, he can choose his investment weights such that diversification is at least partially restored. However, this optimal linear combination of mean-variance efficient portfolios within each asset class usually still differs from the optimally diversified portfolio over all assets. To restore diversification further, Sharpe (1981) suggests that the CIO impose investment rules on one or both of the investment managers to solve an optimization problem that includes the covariances between assets in different asset classes. Elton and Gruber (2004) show that it is possible to overcome the loss of diversification by providing the asset managers with investment rules that they are required to implement. The asset managers can then implement the CIO's optimal strategy without giving up their private information.

Both investment rules described above interfere with the asset manager's desire to maximize his individual performance, on which his compensation depends. Furthermore, when the investment choices of the managers are not always fully observable, these ad hoc rules are not enforceable. In contrast, we propose to change managers' incentives by introducing

a return benchmark against which the managers are evaluated for the purpose of their compensation. When this benchmark is implemented in the right way, it is in the managers' own interest to follow investment strategies that are (more) in line with the objectives of the CIO. In Section 2, we assume that investment opportunities are constant. This allows us to focus on the loss of diversification and on differences in preferences in a parsimonious framework. We then add market-timing skill and horizon effects in Section 3. Both sections assume that the CIO can infer the managers' risk attitudes. This assumption is relaxed in Section 4.

Perhaps one of the most interesting questions is why the CIO should hire multiple asset managers to begin with. Sharpe (1981) argues that the decision to employ multiple managers may be motivated by the desire to exploit their specialization or to diversify among asset managers. Alternatively, Barry and Starks (1984) argue that risk sharing considerations may be a motivation to employ more than one manager. In Section 3, investment opportunities are time-varying, consistent with the empirical evidence that equity and bond returns are to some extent predictable.[2] This allows skilled managers to implement active strategies that generate alphas, when compared to unconditional (passive) return benchmarks. This specific interpretation of alpha may seem unconventional, but it avoids the question of whether asset managers do or do not have private information. Treynor and Black (1973), Admati and Pfleiderer (1997), and Elton and Gruber (2004) assume that managers can generate alpha, but do not explicitly model how managers do so.Cvitanic, Lazrak, Martellini, and Zapatero (2006) assume that the investor is uncertain about the alpha of the manager and derive the optimal policy in that case. We explicitly model the time-variation in investment opportunities and assume that the resulting predictability can be exploited by skilled managers to generate value.

Apart from the tactical aspect of return predictability, time-variation in risk premia can also have important strategic consequences. After all, when asset returns are predictable, the optimal portfolio choice of the CIO depends on his investment horizon.[3] This then requires dynamic optimization to find the optimal composition of the CIO's portfolio. The resulting portfolio choice is referred to as strategic as opposed to myopic (or tactical). The differences between the strategic and myopic portfolio weights are called hedging demands as they hedge against future changes in the investment opportunity set. These hedging demands are usually more pronounced for longer investment horizons of the CIO. As the remuneration schemes of

[2]See, for example, Ang and Bekaert (2007), Lewellen (2004), Campbell and Yogo (2006), Binsbergen and Koijen (2007), and Lettau and van Nieuwerburgh (2006) for stock return predictability, and Dai and Singleton (2002) and Cochrane and Piazzesi (2005) for bond return predictability.

[3]See, for instance, Jorion (1996), Campbell and Viceira (1999), Brandt (1999), Aït-Sahalia and Brandt (2001), Campbell, Chan, and Viceira (2003), and Jurek and Viceira (2007), and Sangvinatsos and Wachter (2005).

investment managers are generally based on a relatively short period, their portfolio weights will be virtually myopic. The CIO, in contrast, usually has a long-term investment horizon. This leads to a third misalignment of incentives.

When unconditional benchmarks are used to overcome costs induced by differences in investment horizons, a key question is whether (i) the benchmark and/or (ii) the strategic allocation to the different asset classes exhibit horizon effects. Most strategic asset allocation papers take a centralized perspective and assume that the tactical and strategic aspects are in perfect harmony.[4] Once investment management is decentralized, tactical and strategic motives are split between the managers and the CIO, respectively. We show that both the strategic allocation, that is, the allocation to the various asset classes, and the optimal benchmarks exhibit strong horizon effects. When investment managers are not constrained by a benchmark, the horizon effects in the strategic allocation are less pronounced, implying that the strategic allocation and optimal benchmarks should be designed jointly.

Our paper also relates to the standard principal-agent literature in which the agent's effort is unobservable. In the delegated portfolio management context, the agent should exert effort to gather the information needed to make the right portfolio decisions, as explored by Ou-Yang (2003).[5] We abstract from explicitly modeling the effort choices of the asset managers. Instead, the managers add value by timing the market, which we assume the CIO cannot do. The agency problem arises because the investment managers, whose actions are not always fully observable, wish to maximize their annual compensation, whereas the CIO acts in the best interest of the beneficiaries of the firm. When designing the benchmarks, the CIO faces a trade-off between (i) allowing the investment managers to realize the gains from market timing and (ii) correcting the misalignments of incentives described above. As a result, the investment problem we solve is nontrivially more difficult than the problem with a CIO and a single investment manager. The strategic allocation of the CIO results from a joint optimization over the benchmark and the strategic allocation to the asset managers.

In the principal-agent literature above, it is common practice to assume that the preferences of the agents (the investment managers) are known to the principal (the CIO). We extend this literature by also considering the realistic case in which the principal has limited knowledge about the agents' preferences. As mentioned before, we assume that the CIO knows the cross-sectional distribution of investment managers' risk appetites, but does not know where in this distribution a given manager falls. We derive (approximate) closed-form solutions for the strategic allocation to the asset classes. In particular, we show that uncertainty about the managers' risk attitudes propagates as a form of background risk (Gollier

[4]Consider, for instance, Campbell, Chan, and Viceira (2003) and Jurek and Viceira (2007).

[5]Stracca (2006) provides a recent survey of the theoretical literature on delegated portfolio management.

and Pratt (1996)), which effectively increases the risk aversion of the CIO. Alternatively, limited knowledge of the managers' risk attitudes can be interpreted as a form of Bayesian parameter uncertainty (see, for example, Barberis (2000), Brennan and Xia (2001)). For ease of exposition, we confine attention to a tractable CRRA preference structure and a realistic linear class of performance benchmarks that are assumed to satisfy the participation constraint of the asset managers.

Finally, our work relates to the organizational literature of Dessein, Garicano, and Gertner (2005), who investigate a general manager (the CIO) who attempts to achieve a common goal while providing strong performance-linked compensation schemes to specialists (the investment managers) to overcome the moral hazard problem. They show that to achieve the common goal, individual incentives may have to be weakened. A common way to align incentives is to give the managers a share in each other's output. Our results indicate that in the portfolio management setting, cross-benchmarking, where the benchmark of an asset manager includes assets from other classes, is not required.[6]

The paper proceeds as follows. In Section 2, we present the model in a financial market with constant investment opportunities. Section 3 extends the financial market by allowing for time-variation in expected returns. In Section 4, we generalize our framework by considering the problem of a CIO who is uncertain about the managers' risk attitudes. Section 5 concludes.

3.2 Constant Investment Opportunities

3.2.1 Financial Market and Preferences

We assume that the financial market contains $2k+1$ assets with prices denoted by S_i, $i = 0, \ldots, 2k$. The first asset, S_0, is a riskless cash account, that evolves according to:

$$\frac{\mathrm{d}S_{0t}}{S_{0t}} = r\mathrm{d}t, \tag{3.1}$$

where r denotes the (constant) instantaneous short rate. The remaining $2k$ assets are risky. We assume that the dynamics of the risky assets are given by geometric Brownian motions. For $i = 1, \ldots, 2k$, we have

$$\frac{\mathrm{d}S_{it}}{S_{it}} = (r + \sigma_i'\Lambda)\,\mathrm{d}t + \sigma_i'\mathrm{d}Z_t, \tag{3.2}$$

[6]For a treatment of decentralized information processing within the firm see Vayanos (2003).

where Λ denotes a $2k$-dimensional vector of, for now, constant prices of risk and Z is a $2k$-dimensional vector of independent standard Brownian shocks. All correlations between asset returns are captured by the volatility vectors σ_i. The volatility matrix of the first k assets is given by $\Sigma_1 = (\sigma_1, \ldots, \sigma_k)'$ and for the second k assets by $\Sigma_2 = (\sigma_{k+1}, \ldots, \sigma_{2k})'$.

The CIO, who acts in the best interest of the beneficiaries of the firm, employs two asset managers. The managers independently decide on the optimal composition of their portfolios using a subset of the available assets. The first asset manager has the mandate to manage the first k assets and the second manager has the mandate to invest in the remaining k assets.

We explicitly model the preferences of both the CIO and the investment managers. Initially, the preference structures are assumed to be common knowledge. We assume that the preferences of the CIO and of the two asset managers can be represented by a CRRA utility function, so that each solves the problem

$$\max_{(x_{is})_{s\in[t,T_i]}} \mathbb{E}_t \left(\frac{1}{1-\gamma_i} W_{T_i}^{1-\gamma_i} \right), \tag{3.3}$$

where γ_i denotes the coefficient of relative risk aversion, T_i denotes the investment horizon, and $i = 1, 2, C$ refers to the two asset managers and the CIO, respectively. The vector x_i denotes the optimal portfolio weights in the different assets available to agent i. According to equation (3), the preferences of the CIO and the investment managers may be conflicting along two dimensions. First, the risk attitudes are likely to be mismatched. Second, the investment horizon used in determining the optimal portfolio choices are potentially different. The remuneration schemes of asset managers usually induce short, say annual, investment horizons. This form of managerial myopia tends to be at odds with the more long-term perspective of the CIO. The difference in horizons is particularly important for CIOs with long-term mandates from pension funds and life insurers.

For now, we assume that investment opportunities are constant. Section 2.1 solves for the optimal portfolio choice when investment management is centralized, implying that the CIO optimizes over the complete asset menu. Obviously, in this case, all misalignments of incentives mentioned before are absent. However, when the investment management firm has a rich investment opportunity set and a substantial amount of funds under management, centralized investment management becomes infeasible. In Section 2.3, we introduce asset managers for each asset class assuming that the asset managers are not constrained by a benchmark. In Section 2.4, the asset managers are then evaluated relative to a performance benchmark, and we show how to design this benchmark optimally. The derivation of the main results is provided in Appendices A to C.

3.2.2 Centralized Problem

As a point of reference, we consider first the centralized problem in which the CIO decides on the optimal weights in all $2k+1$ assets. The instantaneous volatility matrix of the risky assets is given by $\Sigma = (\Sigma_1', \Sigma_2')'$. The corresponding optimal portfolio is given by

$$x_C = \frac{1}{\gamma_C} (\Sigma\Sigma')^{-1} \Sigma\Lambda, \tag{3.4}$$

with the remainder, $1 - x_C'\iota$, invested in the cash account. The utility derived by the CIO from implementing this optimal allocation is

$$J_1(W, \tau_C) = \frac{1}{1-\gamma_C} W^{1-\gamma_C} \exp(a_1 \tau_C), \tag{3.5}$$

where $\tau_C = T_C - t$ and $a_1 = (1-\gamma_C)r + \frac{1-\gamma_C}{2\gamma_C} \Lambda'\Sigma' (\Sigma\Sigma')^{-1} \Sigma\Lambda$. When investment opportunities are constant, the CIO's optimal allocation is independent of the investment horizon, as shown by Merton (1969, 1971).

Suppose that the asset set contains six risky assets. The first three risky assets are fixed income portfolios, namely, a government bond index and two Lehman corporate bond indices with Aaa and Baa ratings, respectively. The remaining three risky assets are equity portfolios made up of firms sorted into value, intermediate, and growth categories based on their book-to-market ratio. The model is estimated by maximum likelihood using data from December 1973 through November 2004. The nominal short rate is set to 5% per annum. Finally, to ensure statistical identification of the elements of the volatility matrix, we assume that Σ is lower triangular.

The estimation results are provided in Table 1. Panel A shows estimates of the parameters Λ and Σ. Panel B shows the implied instantaneous expected return and correlations between the assets. In the fixed income asset class, we find an expected return spread of 1% between corporate bonds with a Baa versus Aaa rating. In the equities asset class, we estimate a high value premium of 4.8%. The correlations within asset classes are high, between 80% and 90%. Furthermore, there is clear dependence between asset classes, which, as we show more formally later, implies that the two-stage investment process leads to inefficiencies.

3.2.3 Decentralized Problem without a Benchmark

We now solve the decentralized problem in which the first asset manager has the mandate to decide on the first k assets and the second asset manager manages the remaining k assets. Neither of the asset managers has access to a cash account. If they did, they could hold highly

leveraged positions or large cash balances, which is undesirable from the CIO's perspective.[7] The CIO allocates capital to the two asset managers and invests the remainder, if any, in the cash account.

The optimal portfolio of asset manager i when he is not constrained by a benchmark is

$$x_i^{NB} = \frac{1}{\gamma_i} x_i + \left(1 - \frac{x_i' \iota}{\gamma_i}\right) x_i^{MV}, \tag{3.6}$$

where

$$x_i = \left(\Sigma_i \Sigma_i'\right)^{-1} \Sigma_i \Lambda \quad \text{and} \quad x_i^{MV} = \frac{\left(\Sigma_i \Sigma_i'\right)^{-1} \iota}{\iota' \left(\Sigma_i \Sigma_i'\right)^{-1} \iota}. \tag{3.7}$$

The optimal portfolio of the asset managers can be decomposed into two components. The first component, x_i, is the standard myopic demand that optimally exploits the risk-return trade-off. The second component, x_i^{MV}, minimizes the instantaneous return variance and is therefore labeled the minimum-variance portfolio. The minimum variance portfolio substitutes for the riskless asset in the optimal portfolio of the asset manager. The two components are then weighted by the risk attitude of the asset manager to arrive at the optimal portfolio.

The CIO has to decide how to allocate capital to the two asset managers as well as to the cash account. We call this decision the strategic asset allocation. The investment problem of the CIO is of the same form as in the centralized problem, but with a reduced asset set. In the centralized setting the CIO has access to $2k+1$ assets. In the decentralized case, each asset manager combines the k assets in his class to form his preferred portfolio. The CIO can then only choose between these two portfolios and the cash account. The instantaneous volatility matrix of the two risky portfolios available to the CIO is given by $\bar{\Sigma} = \left(\Sigma_1' x_1^{NB}, \Sigma_2' x_2^{NB}\right)'$. Thus, the optimal strategic allocation of the CIO to the two asset managers is

$$x_C = \frac{1}{\gamma_C} \left(\bar{\Sigma}\bar{\Sigma}'\right)^{-1} \bar{\Sigma}\Lambda, \tag{3.8}$$

with the remainder, $1 - x_C' \iota$, invested in the cash account. Note that in this case x_C is a two-dimensional vector, containing the strategic allocation to both managers, as opposed to a $2k$-dimensional vector with the weights allocated to each of the assets as in equation (4).

Throughout the paper, utility costs of decentralized investment management are calculated at the centralized level. In other words, we use the value function of the CIO (the principal) to measure utility losses.

[7] A similar cash constraint has been imposed in investment problems with a CIO and a single investment manager (e.g., Brennan (1993) and Gomez and Zapatero (2003)).

The value function of the CIO with decentralization is given by

$$J_2(W, \tau_C) = \frac{1}{1-\gamma_C} W^{1-\gamma_C} \exp(a_2 \tau_C), \tag{3.9}$$

where $\tau_C = T_C - t$ and $a_2 = (1-\gamma_C)r + \frac{1-\gamma_C}{2\gamma_C} \Lambda' \bar{\Sigma}' \left(\bar{\Sigma}\bar{\Sigma}'\right)^{-1} \bar{\Sigma}\Lambda$. It is straightforward to show that the value function in equation (5) (the centralized problem) is larger than or equal to the value function in equation (9) (the decentralized problem). This follows from the fact that the two-stage asset allocation procedure reduces the asset set of the CIO. The CIO can only allocate funds between the two managers, which does not provide sufficient flexibility to always achieve the first-best solution.

The two-stage asset allocation results in the first-best outcome only when the asset managers already happen to implement the proper relative weights within their asset classes. In this case, the CIO can use the strategic allocation to scale up the asset managers' weights to the optimal firm-level allocation. A set of sufficient conditions for this to hold is given by

$$\Sigma_1 \Sigma_2' = 0_{k \times k} \tag{3.10}$$

$$x_i' \iota = \gamma_i, \tag{3.11}$$

with $i = 1, 2$. Note that even when asset classes are independent, that is, Condition (10) holds, the first-best allocation is generally not attainable. If asset classes are independent and when managers do not have access to a cash account, managers allocate their funds to the efficient tangency portfolio and the inefficient minimum-variance portfolio of their asset classes. Condition (11) ensures that the investment in the minimum-variance portfolio equals zero. If both conditions are satisfied, the CIO's optimal strategic allocation to the managers is given by γ_i/γ_C, $i = 1, 2$.

Figure 1 illustrates the solution of the decentralized portfolio problem for a CIO who hires two investment managers with equal risk aversion of 10. Panel A shows the mean-variance (MV) frontier of the bond manager, the MV frontier of the stock manager, and the CIO's optimal linear combination of these two frontiers. The decentralized MV frontier crosses the MV frontier for stocks at the preferred portfolio of the stock manager, and it crosses the MV for bonds at the portfolio chosen by the bond manager. Panel B compares the decentralized MV frontier with the centralized MV frontier. As argued above, the decentralized MV frontier lies within the centralized MV frontier. The welfare loss due to decentralized investment management can be inferred from the difference in Sharpe ratios (i.e., the slope of the lines in MV space through the point $(0, r)$ and tangent to the centralized and decentralized MV frontier, respectively) are also depicted. Finally, panel B also displays the portfolio choices of the CIO for both the centralized and decentralized scenarios. The results clearly show

that the CIO invests more conservatively in the decentralized case. In fact, it can be shown in general that the optimal decentralized portfolio is more conservative than the optimal centralized portfolio.

In Figure 2, we show the utility losses induced by decentralized investment management for various combinations of managerial risk attitudes. The coefficient of relative risk aversion for the CIO equals $\gamma_C = 5$ in Panel A and $\gamma_C = 10$ in Panel B. We define the utility loss as the decrease in the annualized certainty-equivalent return at the firm level. Interestingly, this loss is not minimized when the risk aversion of the asset managers is equal to that of the CIO. In fact, the cost of decentralized investment management is minimized for a risk aversion of 3.3 for the stock manager and 5.7 for the bond manager, regardless of the risk aversion of the CIO. Even though the location of the minimum is not dependent on the risk aversion of the CIO, the utility loss incurred obviously is. When the risk aversion of the CIO equals five, the minimum diversification losses are eight basis points per year in terms of certainty equivalents. This number drops to four basis points when the risk aversion of the CIO equals 10 because he moves out of risky assets and into the riskless asset. The utility loss can increase to 80 to 100 basis points even in this simple example for different risk attitudes of the investment managers. Finally, note that when the CIO is forced to hire a bond manager who does not have the optimal risk aversion level, this may influence the CIO's preferred choice of stock manager and vice versa.

Figure 3 displays the portfolio compositions of the bond manager in Panel A and of the stock manager in Panel B as functions of their risk aversion. Recall that the managers do not have access to a riskless asset. Figure 4 shows the fraction of total risky assets that is allocated to the stock manager as a function of his (and the bond manager's) risk aversion. The bond manager receives one minus this allocation. The allocation of capital between the riskless and the risky assets depends on the risk aversion of the CIO and is not shown.

3.2.4 Decentralized Problem with a Benchmark

We now consider the decentralized investment problem in which the CIO designs a performance benchmark for each of the investment managers in an attempt to align incentives. We restrict attention to benchmarks in the form of portfolios that can be replicated by the asset managers. This restriction implies that only the assets of the particular asset class are used and that the benchmark contains no cash position. There is no possibility and, as we show later, no need for cross-benchmarking. We denote the value of the benchmark of manager i at time t by B_{it} and the weights in the benchmark portfolio for asset class i by β_i. The

evolution of benchmark i is given by

$$\frac{\mathrm{d}B_{it}}{B_{it}} = (r + \beta_i'\Sigma_i\Lambda)\,\mathrm{d}t + \beta_i'\Sigma_i\mathrm{d}Z_t, \tag{3.12}$$

where $\beta_i'\iota = 1$ for $i = 1, 2$.

We assume that the asset managers derive utility from the ratio of the value of assets under their control to the value of the benchmark. They face the problem

$$\max_{(x_{is})_{s\in[t,T_i]}} \mathbb{E}_t\left(\frac{1}{1-\gamma_i}\left(\frac{W_{iT_i}}{B_{iT_i}}\right)^{1-\gamma_i}\right). \tag{3.13}$$

This preference structure can be motivated in several ways. First, the remuneration schemes of asset managers usually contain a component that depends on their performance relative to a benchmark. This is captured in our model by specifying preferences over the ratio of funds under management to the value of the benchmark, in line with Browne (1999) and Browne (2000). Second, investment managers often operate under risk constraints. An important way to measure risk attributable to manager i is to employ tracking error volatility. The tracking error is usually defined as the return differential of the funds under management and the benchmark. Taking logs of the ratio of wealth to the benchmark provides the tracking error in log returns. Third, for investment management firms that need to account for liabilities, such as pension funds and life insurers, supervisory bodies often summarize the financial position by the ratio of assets to liabilities, the so-called funding ratio as further described in Sharpe (2002) and Binsbergen and Brandt (2007). Hence, the ratio of wealth to the benchmark (liabilities) can be interpreted as a reasonable summary statistic of relative performance.[8]

When the performance of asset manager i is measured relative to the benchmark, his optimal portfolio is given by

$$x_i^B = \frac{1}{\gamma_i}x_i + \left(1 - \frac{1}{\gamma_i}\right)\beta_i + \frac{1}{\gamma_i}\left(1 - x_i'\iota\right)x_i^{MV}, \tag{3.14}$$

where x_i and x_i^{MV} are given in equation (7). This portfolio differs from the optimal portfolio in the absence of a benchmark in two important respects. First, the optimal portfolio contains

[8]In addition, Stutzer (2003b) and Foster and Stutzer (2003) show that when the optimal portfolio is chosen so that the probability of underperformance tends to zero as the investment horizon goes to infinity, the portfolio that maximizes the probability decay rate solves a criterium similar to power utility with two main modifications. First, the investor's preferences involve the ratio of wealth over the benchmark. Second, the investor's coefficient of relative risk aversion depends on the investment opportunity set. This provides an alternative interpretation of preferences over the ratio of wealth to the benchmark as well as different coefficients of relative risk aversion for the various asset classes.

a component that replicates the composition of the benchmark portfolio. It is exactly this response of the investment manager that allows the CIO to optimally design a benchmark to align incentives. Note that the benchmark weights enter the optimal portfolio linearly. Second, when the coefficient of relative risk aversion, γ_i, tends to infinity, the asset manager tracks the benchmark exactly. Hence, the benchmark is considered to be the riskless asset from the perspective of the asset manager.

The CIO has to optimally design the two benchmark portfolios and has to determine the allocation to the two asset managers as well as to the cash account. It is important to note that $x_i^B = x_i^{NB}$ when $\beta_i = x_i^{MV}$. That is, the optimal portfolios with and without a performance benchmark coincide when the benchmark portfolio equals the minimum-variance portfolio. This implies that when designing a benchmark, the no-benchmark case is in the choice set of the CIO. As a consequence, the optimal benchmark will reduce the utility costs of decentralized investment management. More importantly, when investment opportunities are constant, the benchmark can be designed so that all inefficiencies are eliminated. The composition of the optimal benchmark that leads to the optimal allocation of the centralized investment problem is given by

$$\beta_i = x_i^{MV} + \frac{\gamma_i}{\gamma_i - 1}\left(\frac{x_i^C}{x_i^{C\prime}\iota} - x_i^{NB}\right), \tag{3.15}$$

where x_i^C are the optimal weights for the assets under management by manager i when the CIO controls all assets as given in equation (4) and x_i^{NB} is given in equation (6). The benchmark weights sum to one because of the restriction that the benchmark cannot contain a cash position.

The two components of the optimal benchmark portfolio have a natural interpretation. The first component is the minimum-variance portfolio. As we point out above, once the benchmark portfolio coincides with the minimum-variance portfolio, the benchmark does not affect the manager's optimal portfolio. The second component, however, corrects the manager's portfolio choice to align incentives. If the relative weights of the CIO and the portfolio of the manager without a benchmark (i.e., x_i^{NB}) coincide, there is no need to influence the manager's portfolio and the second term is zero. However, when the CIO optimally allocates a larger share of capital to a particular asset in class i, the optimal benchmark will contain a positive position in this asset when $\gamma_i > 1$. The ratio before the second component accounts for the manager's preferences. If the manager is more aggressive (i.e., $\gamma_i \to 1$), the benchmark weights are more extreme as the manager is less sensitive to benchmark deviations. If the investor becomes more conservative (i.e., $\gamma_i \to \infty$), we get $x_i^{NB} = x_i^{MV}$ and the benchmark coincides with the relative weights of the CIO.

Finally, the CIO uses the strategic allocation to the two asset managers to implement the optimal firm-level allocation. The optimal weight allocated to each manager is given by $x_i^{C\prime}\iota$, with $i = 1, 2$, and the remainder, $1 - x_1^{C\prime}\iota - x_2^{C\prime}\iota$, is invested in the cash account.

Figure 5 shows the composition of the optimal benchmarks for the bond manager in Panel A and for the stock manager in Panel B as functions of their risk aversion. The mechanism through which the benchmark aligns incentives is particularly clear for the fixed income asset class. Without a benchmark, the bond manager invests too aggressively in corporate bonds with a Baa rating. The optimal benchmark therefore contains a large short position in the same asset that reduces the manager's allocation to Baa-rated bonds. For Aaa-rated bonds, the benchmark provides exactly the opposite incentive.

3.3 Time-varying Investment Opportunities

3.3.1 Financial Market

In Section 2, investment opportunities are constant through time and there are only two inefficiencies caused by decentralized investment management, namely, loss of diversification between asset classes and misalignments in risk attitudes. However, the role of asset managers is rather limited in that they add no value in the form of stock selection or market timing. In this section, we allow investment opportunities, and in particular expected returns, to be time varying and driven by a set of common forecasting variables. This setting allows asset managers to implement active strategies that optimally exploit changes in investment opportunities in their respective asset classes. These active strategies can generate alphas when compared to an unconditional (passive) performance benchmark. Thus, active asset management can be value-enhancing.

This extension of the problem adds several new interesting dimensions to the decentralized investment management problem. First, differences in investment horizons create another misalignment of incentives. The CIO generally acts in the long-term interest of the investment management firm, while asset managers tend to be more shortsighted, possibly induced by their remuneration schemes. When the predictor variables are correlated with returns, it is optimal to hedge future time-variation in investment opportunities.[9] As a consequence, the myopic portfolios held by the asset managers will generally not coincide with the CIO's optimal portfolio that incorporates long-term hedging demands. Second, when a common set of predictor variables affects the investment opportunities in both asset classes, active strategies are potentially correlated. This implies that even if instantaneous returns

[9]See, for instance, Jorion (1996), Campbell and Viceira (1999), Brandt (1999), and Liu (2007).

are uncorrelated, long-term returns can be correlated, which aggravates the loss of diversification due to decentralization. Third, the role of benchmarks is markedly different compared to the case of constant investment opportunities. For the sake of realism, we restrict attention to passive (unconditional) strategies as return benchmarks. As we discussed earlier, Admati and Pfleiderer (1997)show that when the asset manager has private information, an unconditional benchmark can be very costly. After all, the asset managers base their decision on the conditional return distribution, whereas the CIO designs the benchmark using the unconditional return distribution.[10] In their framework, it follows therefore, that unless the benchmark is set equal to the minimum-variance portfolio, it induces a potentially large efficiency loss. In our model, in contrast, the benchmark is used to align incentives in a decentralized investment management firm.

We now consider a more general financial market in which the prices of risk, Λ, can vary over time. More explicitly, we model

$$\Lambda(X) = \Lambda_0 + \Lambda_1 X, \tag{3.16}$$

where X denotes an m-dimensional vector of de-meaned state variables that capture time-variation in expected returns. Although the state variables are time-varying, we drop the subscript t for notational convenience. All portfolios in this section are indexed with either the state realization, X, or the investment horizon, τ, in order to emphasize the conditioning information used to construct the portfolio policies.

Most predictor variables used in the literature, such as term structure variables and financial ratios, are highly persistent. In order to accommodate first-order autocorrelation in predictors, we model their dynamics as Ornstein-Uhlenbeck processes:

$$\mathrm{d}X_{it} = -\kappa_i X_{it}\mathrm{d}t + \sigma'_{Xi}\mathrm{d}Z_t, \tag{3.17}$$

where Z now denotes a $(2k + m)$-dimensional Brownian motion. The volatility matrix of the m predictors is given by $\Sigma_X = (\sigma_{X1}, \ldots, \sigma_{Xm})'$. We assume again that only the CIO has access to a cash account. Finally, we postulate the same preference structures for the CIO and the asset managers as in Section 2.1.

We estimate the return dynamics using three predictor variables: the short rate, the yield on a 10-year nominal government bond, and the log dividend yield of the equity index. These predictors have been used in strategic asset allocation problems to capture the time-variation

[10] Although the predictors are publicly observed, we assume that the CIO is time-constrained or not sufficiently specialized to exploit this information. As such, the conditional return distribution remains unknown to the CIO and the conditioning information exploited by the asset managers is equivalent to private information.

in expected returns (see the references in footnote 3). The model is estimated by maximum likelihood using data from January 1973 through November 2004. The estimation results are presented in Table 2.

The estimates of the unconditional instantaneous expected returns, Λ_0, are similar to the results in Table 1. The second part of Table 2 describes the responses of the expected returns of the individual assets to changes in the state variables, $\Sigma\Lambda_1$. We find that the short rate has a negative impact on the expected returns of all assets except for government bonds. Furthermore, the expected returns of assets in the fixed income class are positively related to the long-term yield, while the expected returns of assets in the equity class are negatively related to this predictor. The dividend yield is positively related to the expected returns of all assets. The estimates of the autoregressive parameters, κ_i, reflect the high persistence of the predictor variables. Finally, the last part of Table 2 provides the joint volatility matrix of the assets and the predictor variables.

3.3.2 Centralized Problem

We first solve again the centralized investment problem in which the CIO manages all assets. This solution serves as a point of reference for the case in which investment management is decentralized. The centralized investment problem with affine prices of risk has been solved by, among others, Liu (2007) and Sangvinatsos and Wachter (2005). We denote the CIO's investment horizon by τ_C. The optimal allocation to the different assets is given by

$$\begin{aligned} x_C\left(X,\tau_C\right) = \frac{1}{\gamma_C}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Lambda(X) + ... \\ \frac{1}{\gamma_C}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Sigma_X'\left(B\left(\tau_C\right) + \frac{1}{2}\left(C\left(\tau_C\right) + C\left(\tau_C\right)'\right)X\right), \end{aligned} \tag{3.18}$$

where expressions for $B\left(\tau_C\right)$ and $C\left(\tau_C\right)$, as well as the derivations of the results in this section are provided in Appendix B. The optimal portfolio contains two components. The first component is the conditional myopic demand that optimally exploits the risk-return trade-off provided by the assets. The second component represents the hedging demands that emerge from the CIO's desire to hedge future changes in the investment opportunity set. This second term reflects the long-term perspective of the CIO. The corresponding value function is given by

$$J_1\left(W,X,\tau_C\right) = \frac{1}{1-\gamma_C}W^{1-\gamma_C}\exp\left\{A\left(\tau_C\right) + B\left(\tau_C\right)'X + \frac{1}{2}X'C\left(\tau_C\right)X\right\}, \tag{3.19}$$

with the coefficients A, B, and C provided in Appendix B.

In Figure 6, we illustrate the composition of the optimal portfolio for different investment horizons when the coefficient of relative risk aversion of the CIO equals either $\gamma_C = 5$ in Panel A or $\gamma_C = 10$ in Panel B. Focusing first on the fixed income asset class, we find substantial horizon effects for corporate bonds. At short horizons, the CIO optimally tilts the portfolio towards Baa-rated corporate bonds and shorts Aaa-rated corporate bonds to take advantage of the credit spread. At longer horizons, the fraction invested in Baa-rated bonds increases even further, while the allocation to Aaa-rated corporate bonds decreases. Switching to the results for the equities asset class, we detect a strong value tilt at short horizons due to the high value premium. The optimal portfolio contains a large long position in value stocks and large short position in growth stocks. However, as the investment horizon increases, the value tilt drops, consistent with the results of Jurek and Viceira (2007).[11]

3.3.3 Decentralized Problem without a Benchmark

We now solve the decentralized problem when the CIO cannot use the benchmark to align incentives. In general, the optimal portfolios of the asset managers depend on both the investment horizon and the state of the economy. However, to make the problem more tractable and realistic, we assume that the investment managers are able to time the market and exploit the time-variation in risk premia, but ignore long-term considerations. That is, asset managers implement the conditional myopic strategy

$$x_i^{NB}(X) = \frac{1}{\gamma_i} x_i(X) + \left(1 - \frac{x_i(X)'\iota}{\gamma_i}\right) x_i^{MV}, \tag{3.20}$$

where

$$x_i(X) = \left(\Sigma_i \Sigma_i'\right)^{-1} \Sigma_i \Lambda(X) \quad \text{and} \quad x_i^{MV} = \frac{\left(\Sigma_i \Sigma_i'\right)^{-1} \iota}{\iota' \left(\Sigma_i \Sigma_i'\right)^{-1} \iota}. \tag{3.21}$$

This particular form of myopia can be motivated by the relatively short-sighted compensation schemes of asset managers. Since the average hedging demands for one-year horizons are negligible, we abstract from the managers' hedging motives in this part of the problem.

The CIO does account for the long-term perspective of the firm through the strategic allocation. However, we assume that the CIO implements a strategic allocation that is

[11] This result is also in line with the findings of Campbell and Vuolteenaho (2004), who explain the value premium by decomposing the CAPM beta into a cash flow beta and a discount rate beta. The cash flow component is highly priced but largely unpredictable. The discount rate component demands a lower price of risk but is to some extent predictable. Campbell and Vuolteenaho (2004) show that growth stocks have a large discount rate beta, whereas value stocks have a large cash flow beta. This implies that from a myopic perspective, value stocks are more attractive than growth stocks. However, the predictability of growth stock returns implies that long-term returns on these assets are less risky, making them relatively more attractive.

unconditional, that is, independent of the current state. At each point in time, the allocation to the different asset classes is reset towards a constant-proportions strategic allocation, as opposed to constantly changing the strategic allocation depending on the state. In order to decide on the strategic allocation, the CIO maximizes the unconditional value function

$$\max_{x_C(\tau_C)} \mathbb{E}\left(J_2\left(W, X, \tau_C\right) \mid W\right), \tag{3.22}$$

where J_2 denotes the conditional value function in the decentralized problem above. Obviously, the CIO's horizon, τ_C, influences the choice of the strategic allocation.

To review the setup of this decentralized problem, the asset managers implement active strategies in their asset classes using conditioning information but ignore any long-term considerations. The CIO, in contrast, allocates capital unconditionally to the asset classes, but accounts for the firm's long-term perspective.

In order to determine the unconditional value function, we evaluate first the conditional value function of the CIO, J_2, for any choice of the strategic allocation. In Appendix B, we show that the conditional value function is exponentially quadratic in the state variables:

$$J_2\left(W, X, \tau_C\right) = \frac{W^{1-\gamma_C}}{1-\gamma_C} \exp\left\{\left(A\left(\tau_C, x_C\right) + B\left(\tau_C, x_C\right)' X + \frac{1}{2} X' C\left(\tau_C, x_C\right) X\right\}. \tag{3.23}$$

One aspect of the CIO's problem is particularly interesting. The active strategy implemented by the asset managers, x_i^{NB}, is affine in the predictor variables: $x_i^{NB}(X) = \zeta_{0i}^{NB} + \zeta_{1i}^{NB} X$. As a consequence, the implied wealth dynamics faced by the CIO are given by

$$\frac{\mathrm{d}W_t}{W_t} = \left(r + \sigma_W(X)'\Lambda(X)\right)\mathrm{d}t + \sigma_W(X)'\mathrm{d}Z_t, \tag{3.24}$$

where $\sigma_W(X)' = x_{1C}\left(\zeta_{01}^{NB} + \zeta_{11}^{NB} X\right)' \Sigma_1 + x_{2C}\left(\zeta_{02}^{NB} + \zeta_{12}^{NB} X\right)' \Sigma_2$. Since the asset managers condition their portfolios on the state variables, the CIO has to allocate capital to two assets that exhibit a very particular form of heteroskedasticity. Hence, despite the homoskedastic nature of the financial market, the CIO is confronted with heteroskedastic asset returns in the decentralized investment management problem.

We solve for the optimal strategic asset allocation numerically (see Appendix B for details). In Figure 7, we present the strategic allocation to the fixed income and equities classes for different investment horizons. The preference parameters are set to $\gamma_C = 10$ and $\gamma_1 = \gamma_2 = 5$. The strategic allocation to the asset classes exhibits substantial horizon effects and marginally overweighs equities. Recall that the strategic allocation to asset classes is independent of the state variables, by construction, because it is unconditional.

Figure 8 provides the annualized utility costs from decentralized asset management for different risk attitudes of the investment managers. The investment horizon equals either $T = 1$ year in Panel A or $T = 10$ years in Panel B. The utility costs are large and increasing in the horizon of the CIO. For relatively short investment horizons, the costs closely resemble the case with constant investment opportunities, with an order of about 40 to 80 basis points per annum. In contrast, for longer investment horizons, the utility costs are substantially higher, around 200 to 300 basis points per annum. Note that the risk attitudes of the managers, for which the costs of decentralized investment management are minimized, depend on the CIO's investment horizon.

3.3.4 Decentralized Problem with a Benchmark

We show in Section 2.4 that when investment opportunities are constant, a performance benchmark can be designed to eliminate all inefficiencies induced by decentralized asset management. This section reexamines this issue for the case of time-varying investment opportunities. We restrict attention to unconditional benchmarks, meaning the benchmark portfolio weights are not allowed to depend on the state variables.[12] Unconditional benchmarks have the advantage that they are easy to implement. Moreover, investment managers following an unconditional benchmark do not have to trade excessively, which could be the case with a conditional benchmark. Conditional benchmarks are more flexible and may therefore reduce further or even eliminate the costs of decentralization.

The performance benchmark of asset manager i is given by a k-dimensional vector of unconditional portfolio weights, β_i, with $\beta_i'\iota = 1$. Since the benchmark is chosen unconditionally, asset managers can outperform their benchmark (i.e., generate alpha) by properly incorporating the conditioning information. The benchmark dynamics are

$$\frac{\mathrm{d}B_{it}}{B_{it}} = (r + \beta_i'\Sigma_i\Lambda(X))\,\mathrm{d}t + \beta_i'\Sigma_i\mathrm{d}Z_t. \tag{3.25}$$

To solve for the optimal benchmark, we first determine the optimal response of the asset managers to their benchmarks. The optimal conditional myopic strategy of the investment managers with a benchmark is given by

$$x_i^B(X) = \frac{1}{\gamma_i}x_i(X) + \left(1 - \frac{1}{\gamma_i}\right)\beta_i + \frac{1}{\gamma_i}\left(1 - x_i(X)'\iota\right)x_i^{MV}, \tag{3.26}$$

where $x_i(X)$ and x_i^{MV} as in equation (21). The CIO chooses the (unconditional) benchmarks and determines the (unconditional) strategic allocation to the asset classes by maximizing

[12] See also Cornell and Roll (2005).

the unconditional expectation of the conditional value function,

$$\max_{x_C(\tau_C),\beta_1(\tau_C),\beta_2(\tau_C)} \mathbb{E}\left(J_3\left(W,X,\tau_C\right) \mid W\right). \tag{3.27}$$

The conditional value function, J_3, is again exponentially quadratic in the state variables and the coefficients are provided in Appendix B. Note that both the strategic allocation and the benchmarks are allowed to depend on the CIO's horizon.

We use numerical methods to solve for the optimal benchmarks and allocations to the two asset classes (see Appendix B for details). Panel A of Figure 9 shows the optimal performance benchmarks for different investment horizons of the CIO. The CIO's risk aversion equals 10 and the managers' risk aversion is set to five. At short horizons, or if the CIO behaves myopically, the optimal benchmarks are similar to when investment opportunities are constant. However, the benchmark portfolios exhibit strong horizon effects. For instance, in the equities asset class, the myopic benchmark reinforces the value tilt already present in the equity manager's (myopic) portfolio. The long-run benchmark, in contrast, anticipates the lower risk of growth stocks and provides an incentive to reduce the value tilt. This illustrates how performance benchmarks can be used to incorporate the CIO's long-term perspective in the short-term portfolio choices of asset managers.

Panel B of Figure 9 provides the corresponding strategic allocation to both asset classes for different investment horizons. Recall that when investment opportunities are constant, the centralized allocation is always more risky than the decentralized allocation without a benchmark. When investment opportunities are time varying, we find the initial allocation with a benchmark to be similar to (and even somewhat more conservative than) the allocation without a benchmark. However, for longer investment horizons of the CIO, the optimal strategic allocation of the CIO is tilted substantially towards equities.

Figure 10 presents the utility gains generated by an optimally chosen benchmark. The CIO's coefficient of risk aversion equals 10 and the horizon is set to $T = 1$ year in Panel A and $T = 10$ years in Panel B. For the 1-year horizon, the value added by the benchmark is limited to approximately 20 basis points. However, when the investment horizon increases to 10 years, the benefit of an optimally chosen benchmark increases as the asset managers become less conservative.

We conclude that unconditional performance benchmarks are significantly value enhancing. This extends the results of Admati and Pfleiderer (1997) concerning the role of performance benchmarks in delegated portfolio management problems. In case of multiple asset managers, performance benchmarks can be useful in aligning incentives along at least three dimensions, namely, diversification, preferences, and investment horizons.

3.4 Unknown Risk Appetites of the Managers

In the previous sections, we assume that the CIO is able to observe the managers' risk aversion levels in deciding on the strategic allocation and in constructing the performance benchmarks. In reality, the CIO usually has relatively limited information about the managers' preferences. Even though past performance or current portfolio holdings can be informative about the managers' risk attitude, exact inference is often infeasible.

In this section therefore, we generalize our framework by explicitly modeling the CIO's uncertainty about the managers' preferences. Specifically, we focus on the impact of the unknown risk aversion levels of the asset managers on (i) the strategic allocation to each of the asset classes, (ii) the utility costs of decentralization, and (iii) the value of optimally designed performance benchmarks. We model the CIO's uncertainty with respect to the managers' risk attitudes by assuming that the CIO has a prior distribution over the risk attitudes of the managers. It is important to note that even when the CIO does not wish to implement optimally designed benchmarks, the CIO needs this prior distribution to decide the strategic allocation to each of the asset classes. We then examine the extent to which the implementation of optimal benchmarks is effective in aligning incentives when the CIO can use no more information than his prior beliefs to design the benchmarks.

We assume that the CIO's prior over the managers' coefficient of relative risk aversion is given by a normal distribution truncated between one and 10.[13] More formally, the prior is given by

$$\gamma \sim f(\gamma) = \frac{\exp\left[-\frac{1}{2}\left(\gamma-\mu_\gamma\right)'\Sigma_\gamma^{-1}\left(\gamma-\mu_\gamma\right)\right]}{\int_1^{10}\int_1^{10}\exp\left[-\frac{1}{2}\left(\gamma-\mu_\gamma\right)'\Sigma_\gamma^{-1}\left(\gamma-\mu_\gamma\right)\right]\mathrm{d}\gamma_1\mathrm{d}\gamma_2},\ \gamma\in(1,10)\times(1,10), \quad (3.28)$$

with $\gamma = (\gamma_1, \gamma_2)$. The parameters μ_γ and Σ_γ allow us to vary the average risk appetites of the asset managers as well as the precision.[14] The off-diagonal elements of Σ_γ allow for correlations between the risk attitudes of the managers. Note that when $\Sigma_{\gamma(1,1)}$ and $\Sigma_{\gamma(2,2)}$ tend to infinity, the prior converges to an uninformative uniform prior on the interval $(1, 10)$. Note further that within our model, the CIO could potentially learn about managerial preferences through the volatility matrix of the managers' portfolio returns (Merton (1980)). We consider learning about the managers' preferences to be beyond the scope of this paper,

[13] Increasing the upper bound of this truncated normal distribution to, for example, 15 or 20 does not affect our qualitative results.

[14] Note that the truncated normal distribution is skewed if μ_γ does not equal the average of the upper and lower truncation points. In this case, changing μ_γ affects the precision and, likewise, changing Σ_γ has an impact on the average risk attitude. To analyze the impact of uncertainty about the managers' preferences by varying Σ_γ, we focus our discussion predominantly on a symmetric prior with $\mu_\gamma = 5.5$. The results for alternative, skewed prior distributions are reported for completeness and are qualitatively similar.

however, and we therefore assume that the uncertainty about the managers' preferences is not alleviated or resolved during the course of the investment period.

In order to determine the optimal strategy of the CIO, we integrate out the uncertainty about the managers' risk aversion levels. This results in a strategic asset allocation and performance benchmarks that are robust to a range of preferences of the asset managers. In Section 4.1, we determine the optimal strategic allocation and the costs of decentralization for different priors over the managers' preferences. Next, we examine in Section 4.2 the extent to which optimal performance benchmarks are useful in reducing the utility costs induced by decentralization. Finally, Section 4.3 introduces tracking error volatility constraints, which are often observed in the investment management industry to constrain asset managers.

3.4.1 Decentralized Problem without a Benchmark

We first consider the case in which the asset managers are not remunerated relative to a benchmark. These managers adopt the strategies given in equation (6). The CIO determines the strategic allocation by maximizing

$$\max_{x_C} \mathbb{E}_t \left(\frac{1}{1-\gamma_C} W_{T_C}^{1-\gamma_C} \right), \tag{3.29}$$

where the expectation is taken with respect to both the uncertainty in the financial market and the risk appetites of the asset managers. We can simplify the problem by first conditioning on the managers' risk aversion levels (γ) and then applying the law of iterated expectations:

$$\max_{x_C} \mathbb{E} \left(\mathbb{E}_t \left(\frac{1}{1-\gamma_C} W_{T_C}^{1-\gamma_C} \middle| \gamma \right) \right). \tag{3.30}$$

The inside expectation, conditional on the managers' preferences and possibly the state variables at time t, can be determined in closed-form for any strategic allocation x_C using the arguments in Sections 2 and 3. To develop the main intuition, we focus initially on the case of constant investment opportunities. The conditional expectation is then given by

$$\mathbb{E}_t \left(\frac{1}{1-\gamma_C} W_{T_C}^{1-\gamma_C} \middle| \gamma \right) = \frac{1}{1-\gamma_C} W_t^{1-\gamma_c} \exp(a(x_C, \gamma)\tau_C), \tag{3.31}$$

where $\tau_C = T_C - t$ and $a(x_C, \gamma) = (1-\gamma_C)\left(x_C \bar{\Sigma}(\gamma)\Lambda + r\right) - \frac{\gamma_C(1-\gamma_C)}{2} x_C' \bar{\Sigma}(\gamma)\bar{\Sigma}(\gamma)' x_C$. Given the prior over the managers' risk appetites, it is straightforward to optimize (numerically) over the strategic allocation. Along these lines we can determine (i) the optimal strategic allocation to both asset classes and (ii) the utility costs induced by decentralization for

various prior distributions over the managers' risk aversion levels.

Even though the results in the remainder of this section are determined numerically, we can illustrate the impact of not knowing the managers' preference parameters using an accurate approximation. The CIO's first-order condition with respect to the strategic allocation, x_C, is given by

$$\mathbb{E}\left(\frac{1}{1-\gamma_C}W_t^{1-\gamma_c}\exp(a(x_C,\gamma)\tau_C)\frac{\partial a(x_C,\gamma)}{\partial x_C}\right)=0_{2\times 1}. \tag{3.32}$$

If the term $\exp(a(x_C,\gamma))$ in equation (32) were constant,[15] the optimal strategic allocation would be given by[16]

$$x_C^{\text{approx}}=\frac{1}{\gamma_C}\left(\mathbb{E}\left(\bar{\Sigma}\bar{\Sigma}'\right)\right)^{-1}\mathbb{E}\left(\bar{\Sigma}\right)\Lambda. \tag{3.33}$$

It is straightforward to show that when the risk appetites of the managers are independent, it follows that

$$\mathbb{E}\left(\bar{\Sigma}\bar{\Sigma}'\right)=\mathbb{E}\left(\bar{\Sigma}\right)\mathbb{E}\left(\bar{\Sigma}\right)'+\begin{bmatrix}\text{Var}\left(\frac{1}{\gamma_1}\right) & 0\\ 0 & \text{Var}\left(\frac{1}{\gamma_2}\right)\end{bmatrix}\begin{bmatrix}b_1'\Sigma_1\Sigma_1'b_1 & 0\\ 0 & b_2'\Sigma_2\Sigma_2'b_2\end{bmatrix}, \tag{3.34}$$

where $b_i=x_i-(x_i'\iota)x_i^{MV}$, $i=1,2$. In other words, b_i is a long-short portfolio that is long the speculative portfolio and short the minimum-variance portfolio. We now discuss the last two matrices on the right-hand side of equation (34) in turn. The first matrix shows that the covariance matrix of managed portfolio returns increases as a result of the uncertainty about the managers' preferences. This induces the CIO to reduce the strategic allocation to each of the asset classes. If the uncertainty about the managers' risk attitudes is equal across managers, this effect is symmetric across asset classes. However, the second matrix depends on the properties of the asset class, which implies that even if the CIO has the same information about the managers' risk attitudes, the relative allocations to the asset classes changes as the uncertainty about the managers' risk attitudes increases.

Using the approximation in equation (33), we can approximate the value function as

$$\exp(a(x_C,\gamma)) \;\simeq\; \exp(a(x_C^{\text{approx}},\gamma))\equiv\exp(\tilde{a}(\gamma)). \tag{3.35}$$

[15] This is the case, for instance, if we consider a 0-th order expansion in $\gamma=\mathbb{E}(\gamma)$.
[16] We normalize $\tau_C=1$.

This approximation allows us to solve the first-order condition (32) in closed-form:

$$x_C^\star \simeq \frac{1}{\gamma_C}\left(\mathbb{E}\left(\exp(\tilde{a}(\gamma))\bar{\Sigma}\bar{\Sigma}'\right)\right)^{-1}\mathbb{E}\left(\exp(\tilde{a}(\gamma))\bar{\Sigma}\right)\Lambda, \tag{3.36}$$

which is similar as before except that the covariance matrix and expected returns are weighted by the (scaled) value function of the CIO ($\exp(\tilde{a}(\gamma))$).

In the empirical application, we treat the uncertainty about the risk aversion levels of both managers symmetrically and assume independence: $\mu_{\gamma(1)} = \mu_{\gamma(2)}$ and $\Sigma_\gamma = \sigma_\gamma^2 I$, with I denoting a 2×2 identity matrix. We consider prior distributions with mean parameters $\mu_\gamma = 3.1$, 5.5, and 7.3 and uncertainty parameters $\sigma_\gamma = 0$, 1, 2, 3, and 25. Note that when $\mu_\gamma = 5.5$ the distribution is symmetric as 5.5 is the average of the truncation points 1 and 10. When $\sigma_\gamma = 25$, the CIO effectively has a uniform prior over γ, and the parameter μ_γ has no further impact.

The results are summarized in Tables 3 and 4. In Table 3 we compute the optimal strategic allocations without benchmarks. In Table 4 we report the corresponding costs of decentralized investment management. Each table has three panels, one for constant investment opportunities (Panel A) and two panels for time-varying investment opportunities with the CIO's investment horizon equal to either $T = 1$ (Panel B) or $T = 10$ (Panel C).

We focus our discussion on the prior distribution with $\mu_\gamma = 5.5$, since this distribution is symmetric. The results in Table 3 indicate that an increase in the uncertainty about the managers' risk aversion leads to a decrease in the optimal allocation to both asset classes. This implies that uncertainty about the managers' preferences effectively increases the risk aversion of the CIO. Not knowing the managers' preferences constitutes a form of background risk, which reduces the investor's appetite for financial risk.[17] The results can also be interpreted as a form of Bayesian parameter uncertainty. This intuition can easily be derived from equations (33) to (36). The effect is quantitatively strong, especially for the equity class. If the prior changes from known preferences (no uncertainty) to a uniform prior between one and 10, the CIO reduces the allocation to the equity asset class by 25% to 50% of the total allocation. Finally, to verify the accuracy of our approximation, we also present in Panel A in parentheses the approximate optimal strategic allocation using equation (36). We conclude that our approximation has a very high level of accuracy, lending further credibility to the intuitive insights it offers.

We report in Table 4 the utility costs incurred by the CIO as a result of decentralization for risk aversion parameters of the CIO equal to $\gamma_C = 5$ and $\gamma_C = 10$. The utility costs are annualized and measured in basis points. The costs of decentralized investment management

[17] See, for instance, Gollier and Pratt (1996).

are generally increasing in the uncertainty about the managers' preferences. The impact of this uncertainty on the utility costs is economically significant. In most cases, the costs double when we move from known levels of risk aversion to a uniform prior distribution over the levels of risk aversion. For instance, in Panel B with $\mu_\gamma = 5.5$ and $\gamma_C = 10$, the utility costs increase from 59 to 109 basis points per annum. These results imply that the common, yet unrealistic, assumption that the preferences of the manager (the agent) are known to the CIO (the principal) can grossly understate the problem and have serious consequences for optimal policies, particularly in the case of time-varying investment opportunities and a long investment horizon for the CIO (see Panel C of Table 3).

Note that there are exceptional cases in which the costs of decentralization are slightly decreasing in the uncertainty about the preferences of the managers. If the CIO assigns a high prior probability to high-cost managers to begin with, which is the case when $\mu_\gamma = 7.3$ and σ_γ is low (see, for instance, Figure 8), increasing σ_γ will increase the probability of allocating capital to lower-cost managers. This in turn can lead to a decreasing relationship between the costs of decentralization and the uncertainty about the managers' preferences. However, this effect is quantitatively negligible and up to only one basis point per year.

3.4.2 Decentralized Problem with a Benchmark

We now examine how effective benchmarks are in aligning incentives if the CIO does not know the risk aversion levels of the managers. Table 5 presents the optimal strategic allocation when the asset managers are remunerated relative to optimal performance benchmarks. The main effects are in line with Table 3. The optimal strategic allocation to both asset classes decreases as the uncertainty about the managers' risk appetites increases. We also find that the implementation of optimal benchmarks can lead to either an increase or decrease in the strategic allocation relative to the problem without benchmarks, depending on the CIO's prior beliefs.

In the previous subsection, we argue that the inefficiencies caused by decentralization are generally aggravated when the risk appetites of the managers are unknown (Table 4). The value of an optimally designed benchmark (Table 6) depends on the following two effects. First, compared to the case of known risk appetites, the amount of information that can be used to design the optimal benchmarks is lower because risk appetites are now unknown. This suggests that the value of an optimal benchmark diminishes. Second, the inefficiencies that can potentially be mitigated by the benchmarks are also much larger. Therefore, there is more scope for the benchmarks to have value-added. This explains why, for low levels of uncertainty, there is a (small) negative relation between the value of benchmarks and the level of uncertainty about the risk appetites. In these cases the first effect dominates. However,

as the uncertainty about the risk aversion levels increases, the value of the benchmarks also generally increases and exceeds the value for known preferences because then the second effect dominates.

As explained before, there are exceptional cases in which the costs of decentralization are slightly decreasing in the uncertainty about the managers' preferences (e.g., when $\mu_\gamma = 7.3$). In such cases, increasing the uncertainty about the managers' preferences does not sufficiently enlarge the scope for improvement by optimally designed benchmarks. As a result, the fact that the benchmarks are based on less information dominates and the value of an optimally designed benchmark decreases in the uncertainty about the managers' risk aversion levels.

When the CIO's investment horizon is longer, for example $T = 10$, the results in Panel C may give the impression that benchmarks become less effective in aligning incentives when risk appetites are unknown. It is important to emphasize, however, that the results presented for this case constitute a conservative lower bound on the value of benchmarks. It is common practice in the investment management industry to have the opportunity to revise the benchmark annually. We consider a single, unconditional benchmark that is held constant for 10 years, which is the absolute minimum of what optimally designed benchmarks can actually achieve. In case of annual rebalancing, or an effective 1-year horizon, Panel B shows that the benchmarks are indeed more effective the more uncertain the CIO is about the managers' risk preferences.

To summarize, we find that uncertainty about the managers' risk preferences has a strong effect on the optimal strategic allocation to the different asset classes. We show that this uncertainty increases the costs of decentralized investment management even further. We also show that optimally designed performance benchmarks become more effective to overcome these costs.

3.4.3 Risk Constraints

Apart from designing optimal return benchmarks, the CIO can also employ risk constraints in order to change or restrict the behavior of asset managers. These risk constraints can be formulated either in terms of absolute risk in absence of a benchmark or in terms of relative risk when the asset manager is remunerated relative to a benchmark. Absolute risk constraints restrict the total volatility of the portfolio return. Relative risk constraints limit the volatility of the portfolio return in excess of the benchmark return, as in Roll (1992) and Jorion (2003). We assume that the volatility constraints have to be satisfied at every point in time. In modern investment management firms, risk management systems monitor the risk exposures of portfolio holdings frequently, which makes it plausible to presume that

risk constraints have to be satisfied continuously. For ease of exposition, we focus initially on the financial market of Section 2 in which investment opportunities are constant.

The instantaneous volatility of the portfolio return is given by $\sigma^A(x_i) = \sqrt{x_i'\Sigma_i\Sigma_i'x_i}$, which is the portfolio's absolute risk. The instantaneous volatility of the portfolio return in excess of the benchmark (relative risk) is given by $\sigma^R(x_i) = \sqrt{(x_i-\beta_i)'\Sigma_i\Sigma_i'(x_i-\beta_i)}$, which is also called the tracking error volatility. Using these definitions for absolute and relative risk, we impose risk limits of the form

$$\sigma^j(x_i) \leq \phi_{ij}, \tag{3.37}$$

with $j = A, R$. To ensure that the optimization problem of the asset managers is well defined, we assume that $\sigma^A\left(x_i^{MV}\right) \leq \phi_{iA}$, which states that the limit on absolute risk must exceed the volatility of the minimum variance portfolio. In the case of relative risk constraints, we require that $\phi_{iR} \geq 0$, since we restrict attention to benchmarks that can be replicated by the managers. A relative risk limit of $\phi_{iR} = 0$ implies that the asset manager must exactly implement the benchmark portfolio. We focus on the effect of imposing either one of these constraints, but not both.[18]

Whenever the unconstrained portfolio choice in absence of a benchmark does not violate the absolute risk constraint, this portfolio remains optimal for manager i. However, once the absolute risk constraint is violated, Appendix C shows that the optimal portfolio equals

$$x_i^{NB}(\xi_i) = \frac{1}{\gamma_i(1+\xi_i)}x_i + \left(1 - \frac{x_i'\iota}{\gamma_i(1+\xi_i)}\right)x_i^{MV}, \tag{3.38}$$

where x_i and x^{MV} are given by equation (7) and $\xi_i > 0$ satisfies $\sigma^A\left(x_i^{NB}(\xi_i)\right) = \phi_{Ai}$. This solution shows that the absolute risk constraint induces an effective increase in risk aversion. The results in Figure 2 then imply that absolute risk constraints can mitigate inefficiencies whenever the investment manager is too aggressive. In contrast, when the investment manager is too conservative, absolute risk constraints can actually aggravate the inefficiencies.

We also show in Appendix C that the optimal portfolio in the presence of a performance benchmark and binding relative risk constraints is given by

$$x_i^B(\xi_i) = \frac{1}{\gamma_i(1+\xi_i)}x_i + \left(1 - \frac{1}{\gamma_i(1+\xi_i)}\right)\beta_i + \frac{1-x_i'\iota}{\gamma_i(1+\xi_i)}x_i^{MV}, \tag{3.39}$$

where x_i and x^{MV} are given in equation (7) and $\xi_i > 0$ satisfies $\sigma^R\left(x_i^{NB}(\xi_i)\right) = \phi_{Ri}$. In

[18] Jorion (2003) infers in addition the effect of implementing both absolute and relative risk constraints.

addition, Appendix C shows that the relative risk constraint binds for an investment manager with risk aversion γ_i once the benchmark is designed on the basis of a higher risk aversion $\tilde{\gamma}_i$, with $\tilde{\gamma}_i > \gamma_i$. This implies that the CIO does not require specific knowledge of the manager's risk attitude, more than knowing an upper bound. If the benchmark and relative risk constraint are designed on the basis of this conservative upper bound, the relative risk constraint binds for more aggressive managers. The binding constraint induces an effective increase in the manager's risk aversion to the level for which the benchmark is designed.

Combining these results with our discussion of unknown risk appetites, risk constraints essentially shift the lower truncation point of the CIO's prior over the managers' risk aversion levels upwards. All managers, who are more aggressive than the risk constraint allows, will behave as an asset manager for which the constraint binds marginally. Hence, risk constraints effectively reduce the CIO's uncertainty about the manager's preferences.

In case of constant investment opportunities, there is no disadvantage from selecting tight risk constraints. However, in the more realistic case of time-varying investment opportunities, the same derivation is valid, albeit ξ_i becomes time dependent and the constraint will bind only at certain points in time. In that case, tight risk constraints will reduce the timing ability of the asset managers. Therefore, the CIO can optimally determine the strategic allocation to both asset classes, the benchmarks for each manager, and the risk constraints for a given prior over the managers' risk tolerances. Tight risk constraints indicate that it is valuable for the CIO to reduce uncertainty about the managers' risk attitude, while wide risk constraints indicate that the CIO prefers to exploit the timing expertise of the managers rather than reducing the uncertainty about their preferences.

3.5 Conclusions

We address several misalignments of incentives generated by decentralized investment management. These misalignments between a CIO and the asset managers he employs can lead to large utility costs. One straightforward solution is to implement centralized investment strategies whereby the CIO attempts to manage all assets himself. However, from an organizational point of view, decentralized investment management is an inevitable and stylized fact of the investment industry. We show in this paper that the optimal design of an unconditional linear benchmark can be very effective in mitigating the costs of decentralized investment management. This is even more pronounced when we generalize our model by relaxing the assumption that the CIO knows the risk aversion levels of the asset managers. The optimal benchmark is derived assuming that the CIO only knows the cross-sectional distribution of investment managers' risk appetites, but does not know where in this distribution

a given manager falls.

For ease of exposition, we confine attention to CRRA preferences and linear performance benchmarks. Future work could focus on a more complicated preference structure and/or nonstandard contracts. For example, it seems reasonable that the utility function of the CIO is kinked as in Binsbergen and Brandt (2007). The compensation scheme for the asset managers may also be nonlinear and/or asymmetric, as in Browne (1999), Browne (2000), Carpenter (2000), and Basak, Pavlova, and Shapiro (2007b), for example. Another interesting extension would be to assess the asset pricing implications of decentralized investment management. In delegated portfolio choice problems, Brennan (1993), Gomez and Zapatero (2003), Cuoco and Kaniel (2006), and Cornell and Roll (2005) illustrate the impact of delegation and benchmarking on equilibrium asset prices. Stutzer (2003a) shows that multiple benchmarks imply a factor model with these benchmarks returns as possibly priced factors. Finally, we show that not knowing the risk preferences of the managers to which the CIO delegates the available capital effectively increases the CIO's risk aversion. Since the amount of capital managed institutionally has increased dramatically during recent decades, it is important to further understand the asset pricing implications of unknown risk preferences.

3.A Constant Investment Opportunities

3.A.1 Decentralized Problem with a Benchmark

We solve the decentralized problem with the optimally designed benchmark of Section 2.4. We derive first the optimal allocations of the asset managers in the presence of a benchmark. Define normalized wealth as $w_{it} = W_{it}B_{it}^{-1}$. Recall that the benchmark comprises only positions in the assets available to the investment managers and no cash. The asset managers are therefore able to replicate the benchmark. The dynamics of the benchmark are given in equation (12). Using Ito's lemma, the dynamics of normalized wealth are

$$\frac{\mathrm{d}w_t}{w_t} = \left(x_i^{B\prime}\Sigma_i\Lambda - \beta_i'\Sigma_i\Lambda + \beta_i'\Sigma_i\Sigma_i\beta_i - \beta_i'\Sigma_i\Sigma_i'x_i^B\right)\mathrm{d}t + \left(x_i^{B\prime}\Sigma_i - \beta_i'\Sigma_i\right)\mathrm{d}Z_t. \tag{3.40}$$

The corresponding Hamilton-Jacobi-Bellman (HJB) equation is

$$\max_{x_i^B : x_i^{B\prime}\iota = 1} \left(\begin{array}{l} J_w w \left(x_i^{B\prime}\Sigma_i\Lambda - \beta_i'\Sigma_i\Lambda + \beta_i'\Sigma_i\Sigma_i\beta_i - \beta_i'\Sigma_i\Sigma_i'x_i^B\right) + \\ \frac{1}{2}w^2 J_{ww}\left(x_i^{B\prime}\Sigma_i - \beta_i'\Sigma_i\right)\left(x_i^{B\prime}\Sigma_i - \beta_i'\Sigma_i\right)' + J_t \end{array} \right) = 0. \tag{3.41}$$

The first-order conditions (FOC) are

$$0 = J_w w\left(\Sigma_i\Lambda - \Sigma_i\Sigma_i'\beta_i\right) + J_{ww}w^2\Sigma_i\left(\Sigma_i'x_i^B - \Sigma_i'\beta_i\right) - \xi\iota, \text{ and } 1 = x_i^{B\prime}\iota, \tag{3.42}$$

with ξ denoting the Lagrange multiplier. The value function is of the form

$$J_3(W/B, \tau_i) = \frac{1}{1-\gamma_i}\left(\frac{W}{B}\right)^{1-\gamma_i}\exp(c\tau_i), \tag{3.43}$$

with $\tau_i = T_i - t$. The solution of the FOCs is given by equation (14).

The CIO has to design the benchmarks, that is, β_i, $i = 1, 2$, and decide on the strategic allocation to the managers and to the cash account. Since the managers' optimal portfolios are affine in the benchmark weights, (see equation (14)), the benchmark can be designed to solve for the optimal relative fractions invested in the different assets present in the asset classes. The strategic allocation, $x_C \in \mathbb{R}^2$, can subsequently be used to optimally manage the absolute fractions allocated to the different assets. More formally, the optimal portfolio is given by

$$x_C = \left[\begin{array}{c} x_{1C} \\ x_{2C} \end{array} \right] = \frac{1}{\gamma_C}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Lambda, \tag{3.44}$$

where x_{iC} denotes the allocation to the assets managed by manager i. We use β_i to solve for the optimal relative fractions invested within the asset class:

$$x_i^B = x_{iC}\left(x_{iC}'\iota\right)^{-1}. \tag{3.45}$$

The optimal benchmark weights are given by

$$\beta_i = \frac{\gamma_i}{\gamma_i - 1}\left[x_i^C\left(x_i^{C\prime}\iota\right)^{-1} - \left(\frac{1}{\gamma_i}x_i + \frac{1}{\gamma_i}\left(1 - x_i'\iota\right)x_i^{MV}\right)\right], \tag{3.46}$$

and the optimal allocation of the CIO's wealth to the managers is given by $x_{iC}'\iota$.

3.B Time-varying Investment Opportunities

3.B.1 Centralized Problem

The centralized problem in Section 3.2 relates to the portfolio choice problems in Sangvinatsos and Wachter (2005) and Liu (2007). The problem is solved using standard dynamic programming techniques. The HJB

equation reads

$$\max_{x_C} \left(\begin{array}{l} J_W W \left(r + x_C' \Sigma \Lambda(X)\right) + \frac{1}{2} J_{WW} W^2 x_C' \Sigma\Sigma' x_C + J_t - \\ J_X' K X + \frac{1}{2} tr\left(\Sigma_X' J_{XX} \Sigma_X\right) + W x_C' \Sigma\Sigma_X' J_{WX} \end{array} \right) = 0, \tag{3.47}$$

where we omit the indices of $x_C(X, \tau_C)$ for notational convenience and $K = \text{diag}(\kappa_1, \ldots, \kappa_m)$. The affine structure of the financial market implies that the value function is exponentially quadratic in the state variables:

$$J(W, X, \tau_C) = \frac{W^{1-\gamma_C}}{1-\gamma_C} \exp\left\{ A(\tau_C) + B(\tau_C)' X + \frac{1}{2} X' C(\tau_C) X \right\}. \tag{3.48}$$

Solving for the FOC of problem (B1) and using equation (B2) to determine the partial derivatives, we obtain

$$x_C(X, \tau_C) = \frac{1}{\gamma_C} (\Sigma\Sigma')^{-1} \Sigma \left[\Lambda(X) + \Sigma_X' \left(B(\tau_C) + \frac{1}{2} \left(C(\tau_C) + C(\tau_C)' \right) X \right) \right], \tag{3.49}$$

which we can rewrite as $x_C(X, \tau_C) = \zeta_0^C(\tau_C) + \zeta_1^C(\tau_C) X$, with

$$\zeta_0^C(\tau_C) = \frac{1}{\gamma_C} (\Sigma\Sigma')^{-1} \Sigma \left[\Lambda_0 + \Sigma_X' B(\tau_C) \right], \tag{3.50}$$

$$\zeta_1^C(\tau_C) = \frac{1}{\gamma_C} (\Sigma\Sigma')^{-1} \Sigma \left[\Lambda_1 + \frac{1}{2} \Sigma_X' \left(C(\tau_C) + C(\tau_C)' \right) \right]. \tag{3.51}$$

To find the coefficients A, B, and C, we substitute the optimal portfolio into the HJB equation (B1) and match the constant, the terms linear in X, and the terms quadratic in X. In what follows, we derive the value function for any affine policy, $x(X, \tau) = \zeta_0(\tau) + \zeta_1(\tau) X$, which turns out be useful in subsequent derivations. The value function for this particular problem is obtained for $\zeta_0(\tau) = \zeta_0^C(\tau)$ and $\zeta_1(\tau) = \zeta_1^C(\tau)$. The resulting ODEs are

$$\begin{array}{rcl} \dot{A} & = & (1-\gamma_C)(r + \zeta_0' \Sigma \Lambda_0) - \frac{1}{2}\gamma_C(1-\gamma_C)\zeta_0' \Sigma\Sigma' \zeta_0 + \\ & & \frac{1}{4} tr\left(\Sigma_X' (C + C') \Sigma_X\right) + \frac{1}{2} B' \Sigma_X \Sigma_X' B + (1-\gamma)\zeta_0' \Sigma\Sigma_X' B, \\ \dot{B}' & = & (1-\gamma_C)\left[\zeta_0' \Sigma \Lambda_1 + \Lambda_0' \Sigma' \zeta_1\right] - \gamma_C(1-\gamma_C)\zeta_0' \Sigma\Sigma' \zeta_1 - B' K + \\ & & \frac{1}{2} B' \Sigma_X \Sigma_X' (C + C') + \frac{1}{2}(1-\gamma_C)\zeta_0' \Sigma\Sigma_X' (C + C') + (1-\gamma_C) B' \Sigma_X \Sigma' \zeta_1, \\ \dot{C} & = & 2(1-\gamma_C)\zeta_1' \Sigma \Lambda_1 - \gamma_C(1-\gamma_C)\zeta_1' \Sigma\Sigma' \zeta_1 - (C + C') K + \\ & & \frac{1}{4} (C + C') \Sigma_X \Sigma_X' (C + C') + (1-\gamma_C)\zeta_1' \Sigma\Sigma_X' (C + C'), \end{array} \tag{3.52}$$

subject to the boundary conditions $A(0) = 0$, $B(0) = 0_{m \times 1}$, and $C(0) = 0_{m \times m}$.

3.B.2 Decentralized Problem without a Benchmark

In the decentralized problem without a benchmark in Section 3.3, we first solve for the myopic, cash-constrained policy of the managers. The optimization problem of the (myopic) managers can be simplified to

$$\max_{x_i^{NB} : x_i^{NB\prime} \iota = 1} \mathbb{E}_t \left(x_i^{NB\prime}(X) \Sigma_i \Lambda(X) - \frac{\gamma_i}{2} x_i^{NB\prime}(X) \Sigma_i \Sigma_i' x_i^{NB}(X) \right). \tag{3.53}$$

As a result, the optimal strategy of the myopic, cash-constrained investment managers is

$$x_i^{NB}(X) = \frac{1}{\gamma_i} x_i(X) + \left(1 - \frac{x_i(X)' \iota}{\gamma_i} \right) x_i^{MV} = \zeta_{0i}^{NB} + \zeta_{1i}^{NB} X, \tag{3.54}$$

where $x_i(X)$ and x_i^{MV} are given in equation (21) and

$$\zeta_{0i}^{NB} = \frac{1}{\gamma_i} (\Sigma_i \Sigma_i')^{-1} \Sigma_i \Lambda_0 + x_i^{MV} \left(1 - \frac{\iota' (\Sigma_i \Sigma_i')^{-1} \Sigma_i \Lambda_0}{\gamma_i} \right), \tag{3.55}$$

$$\zeta_{1i}^{NB} = \frac{1}{\gamma_i} (\Sigma_i \Sigma_i')^{-1} \Sigma_i \Lambda_1 - x_i^{MV} \left(\frac{\iota' (\Sigma_i \Sigma_i')^{-1} \Sigma_i \Lambda_1}{\gamma_i} \right). \tag{3.56}$$

Anticipating the allocations of the asset managers, the CIO has to decide on the strategic allocation. We consider strategic allocations that are independent of the current state of the economy, but that do account for the investment horizon of the CIO. We optimize the unconditional value function

$$\mathbb{E}\left(J_2(W,X,\tau_C)\mid W\right), \tag{3.57}$$

with $J_2(W,X,\tau_C)$ denoting the conditional value function, which is exponentially quadratic in the state variables. After all, if we denote the allocation to the ith asset manager by x_{iC}, then the resulting portfolio of the CIO is affine in the state variables:

$$x_C^{\text{Implied}} = \begin{bmatrix} x_{1C}\left(\zeta_{01}^{NB}+\zeta_{11}^{NB}X\right) \\ x_{2C}\left(\zeta_{02}^{NB}+\zeta_{12}^{NB}X\right)\end{bmatrix} = \zeta_0^{\text{Implied}}+\zeta_1^{\text{Implied}}X, \tag{3.58}$$

and the results of Appendix B apply. To determine the unconditional value function, we use Lemma 1.

LEMMA 1: *Let $Y\in\mathbb{R}^{m\times 1}$, $Y\sim N(0,\Sigma)$, $a\in\mathbb{R}^{m\times 1}$, and $B\in\mathbb{R}^{m\times m}$. If $\left(\Sigma^{-1}-2B\right)$ is strictly positive definite, then we have*

$$\mathbb{E}\left(\exp\left(a'Y+Y'BY\right)\right) = \exp\left(-\frac{1}{2}\ln\det\left(I-2\Sigma B\right)+\frac{1}{2}a'\left(\Sigma^{-1}-2B\right)^{-1}a\right). \tag{3.59}$$

Solving for the optimal strategic asset allocation is then reduced to a static optimization of the unconditional value function, which we perform numerically.

3.B.3 Decentralized Problem with a Benchmark

The performance benchmark of manager i in Section 3.4 is parameterized by a vector of constant portfolio weights, β_i, with the corresponding dynamics specified in equation (25). The asset manager is concerned with wealth relative to the value of the benchmark. The dynamics of normalized wealth, $w_{it}=W_{it}B_{it}^{-1}$, are given by

$$\frac{\mathrm{d}w_t}{w_t} = \left(x_i^{B\prime}(X)\Sigma_i\Lambda(X)+\beta_i'\Sigma_i\left[\Sigma_i'\beta_i-\Lambda(X)-\Sigma_i'x_i^B(X)\right]\right)\mathrm{d}t+\left(x_i^{B\prime}(X)\Sigma_i-\beta_i'\Sigma_i\right)\mathrm{d}Z_t,$$

where $x_i^B(X)$ denotes the myopic conditional portfolio choice of investment manager i. We first optimize the managers' portfolios when they have no access to a cash account, that is, $x_i^{B\prime}\iota=1$. The optimal strategy of the managers is given by

$$x_i^B(X) = \tfrac{1}{\gamma_i}x_i(X)+\left(1-\tfrac{1}{\gamma_i}\right)\beta_i+\tfrac{1}{\gamma_i}\left(1-x_i(X)'\iota\right)x_i^{MV} = \zeta_{0i}^B+\zeta_{1i}^BX, \tag{3.60}$$

where $x_i(X)$ and x_i^{MV} as in equation (21) and

$$\zeta_{0i}^B = \frac{1}{\gamma_i}\left(\Sigma_i\Sigma_i'\right)^{-1}\Sigma_i\Lambda_0+\left(1-\frac{1}{\gamma_i}\right)\beta_i+\frac{1}{\gamma_i}x_i^{MV}\left(1-\iota'\left(\Sigma_i\Sigma_i'\right)^{-1}\Sigma_i\Lambda_0\right) \tag{3.61}$$

$$\zeta_{1i}^B = \frac{1}{\gamma_i}\left(\Sigma_i\Sigma_i'\right)^{-1}\Sigma_i\Lambda_1-\frac{1}{\gamma_i}x_i^{MV}\left(\iota'\left(\Sigma_i\Sigma_i'\right)^{-1}\Sigma_i\Lambda_1\right). \tag{3.62}$$

The implication of equation (B14) is that the optimal portfolio of the managers is again affine in the state variables. The CIO selects the optimal constant proportions strategy and the constant benchmarks, β_1 and β_2, to optimize the unconditional value function, that is, equation (B11). This yields

$$x_C^{\text{Implied}} = \begin{bmatrix} x_{1C}\left(\zeta_{01}^{B}+\zeta_{11}^{B}X\right) \\ x_{2C}\left(\zeta_{02}^{B}+\zeta_{12}^{B}X\right)\end{bmatrix} = \zeta_0^{\text{Implied}}+\zeta_1^{\text{Implied}}X, \tag{3.63}$$

where ζ_{0i}^B and ζ_{1i}^B obviously depend on the choice of the benchmark. The conditional value function is exponentially quadratic as in equation (B2), with $\zeta_0(\tau)=\zeta_0^{\text{Implied}}$ and $\zeta_1(\tau)=\zeta_1^{\text{Implied}}$. The coefficients satisfy the ODEs given in equation (B6). To solve for the strategic allocation and the performance benchmark,

we evaluate the unconditional expectation of the conditional value function using Lemma 1. We then optimize numerically.

3.C Risk Constraints

We derive in this section the optimal allocations of the asset managers in the presence of either relative or absolute risk constraints as defined in Section 4.3. We assume that investment opportunities are constant.

For the case with absolute risk constraints, the optimization problem of asset manager i can be simplified to

$$\max_{x_i^{NB}\in\mathcal{A}_i}\left(x_i^{NB\prime}\Sigma_i\Lambda + r - \frac{\gamma}{2}x_i^{NB\prime}\Sigma_i\Sigma_i'x_i^{NB}\right). \tag{3.64}$$

and the set $\mathcal{A}_i$ is given by $\mathcal{A}_i = \left(x \mid x'\iota = 1, \sqrt{x'\Sigma_i\Sigma_i'x} \leq \phi_{Ai}\right)$. Consequently, the Kuhn-Tucker FOCs are

$$\begin{aligned} 0 &= \Sigma_i\Lambda - \gamma(1+\xi_1)\Sigma_i\Sigma_i'x_i^{NB} - \xi_1\iota & (3.65)\\ 1 &= x_i^{NB\prime}\iota,\ \phi_{Ai}^2 \geq x_i^{NB\prime}\Sigma_i\Sigma_i'x_i^{NB},\ \xi_2 \geq 0 & (3.66)\\ 0 &= \xi_2\left(\phi_{Ai}^2 - x_i^{NB\prime}\Sigma_i\Sigma_i'x_i^{NB}\right), & (3.67)\end{aligned}$$

with ξ_1 and ξ_2 denoting the Kuhn-Tucker multipliers. In fact, ξ_2 is the multiplier for the risk constraint scaled by a factor $\gamma/2$ to simplify the interpretation. If the risk constraint is not binding, the managers' optimal portfolio is as derived in Section 2.3. Otherwise, the absolute risk constraint binds and the optimal portfolio is given by the solution to equation (C2) for $\xi_2 > 0$ so that the risk constraint holds with equality. This results immediately in the optimal portfolio given in equation (38).

When the asset managers have to satisfy relative risk constraints, their objective is

$$\max_{x_i^B\in\mathcal{B}_i}\left(x_i^{B\prime}\Sigma_i\Lambda + \beta_i'\Sigma_i\left[\Sigma_i\beta_i - \Lambda - \Sigma_i'x_i^B\right] - \frac{\gamma}{2}\left(x_i^{B\prime}\Sigma_i - \beta_i'\Sigma_i\right)\left(x_i^{B\prime}\Sigma_i - \beta_i'\Sigma_i\right)'\right), \tag{3.68}$$

where the set $\mathcal{B}_i$ is given by $\mathcal{B}_i = \left(x \mid x'\iota = 1, \sqrt{(x-\beta_i)'\Sigma_i\Sigma_i'(x-\beta_i)} \leq \phi_{Ri}\right)$.

The FOCs are given by

$$\begin{aligned} 0 &= \Sigma_i\left(\Lambda - \Sigma_i'\beta_i\right) - \gamma(1+\xi_1)\Sigma_i\Sigma_i'\left(x_i^B - \beta_i\right) - \xi_1\iota, & (3.69)\\ 1 &= x_i^{B\prime}\iota,\ \phi_{Ri}^2 \geq \left(x_i^B - \beta_i\right)'\Sigma_i\Sigma_i'\left(x_i^B - \beta_i\right),\ \xi_2 \geq 0, & (3.70)\\ 0 &= \xi_2\left(\phi_{Ri}^2 - \left(x_i^B - \beta_i\right)'\Sigma_i\Sigma_i'\left(x_i^B - \beta_i\right)\right), & (3.71)\end{aligned}$$

where ξ_1 and ξ_2 indicate the Kuhn-Tucker multipliers. Again, if the relative risk constraint is not binding, the optimal portfolio of Section 2.4 prevails. Otherwise, the optimal strategy of manager i is given by the solution to equation (C6) with $\xi_2 > 0$ so that the relative risk constraint is satisfied with equality. This implies the strategy given in equation (39).

Finally, suppose that the benchmark is designed on the basis of a higher risk aversion level, say $\tilde{\gamma}$, than the manager's risk aversion, denoted by γ. In this case, the (relative) risk of the manager's portfolio will exceed the (relative) risk that would correspond to a manager with risk aversion level $\tilde{\gamma}$. If the risk limit is constructed for a manager with risk aversion $\tilde{\gamma}$, then the relative risk constraint will bind for the manager with risk aversion γ. This induces an effective increase in the manager's risk aversion from γ to $\tilde{\gamma}$. To show this, note that the difference between the optimal portfolio of the manager, who has a risk aversion γ, and the benchmark weights, which are designed for a manager with risk aversion $\tilde{\gamma}$, is given by

$$x^B\left(\gamma, \beta\left(\tilde{\gamma}\right)\right) - \beta\left(\tilde{\gamma}\right) = \frac{\tilde{\gamma}}{\tilde{\gamma}-1}\frac{1}{\gamma}\left\{x_i - x_i^C\left(x_i^{C\prime}\iota\right)^{-1} + \left(1 - \iota'x_i\right)x_i^{MV}\right\}. \tag{3.72}$$

In this expression, $x^B\left(\gamma, \beta(\tilde{\gamma})\right)$ denotes the optimal portfolio choice when the investor has a coefficient of relative risk aversion γ, but is evaluated relative to a benchmark, $\beta\left(\tilde{\gamma}\right)$, which is based on $\tilde{\gamma}$. This immediately

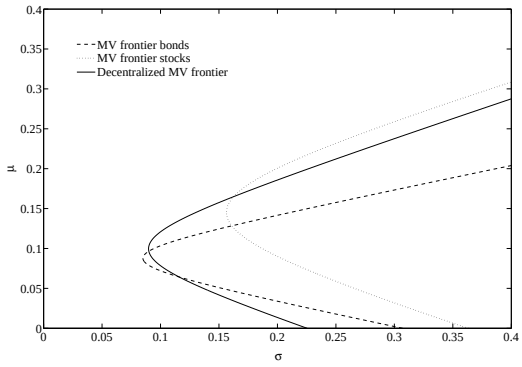
implies for the relative risk of the manager's portfolio:

$$\begin{aligned}\left(x^B\left(\gamma,\beta\left(\tilde{\gamma}\right)\right)-\beta\left(\tilde{\gamma}\right)\right)'\left(\Sigma_i\Sigma_i'\right)\left(x^B\left(\gamma,\beta\left(\tilde{\gamma}\right)\right)-\beta\left(\tilde{\gamma}\right)\right)=\\ \left(\frac{\tilde{\gamma}}{\gamma}\right)^2\left(x^B\left(\tilde{\gamma},\beta\left(\tilde{\gamma}\right)\right)-\beta\left(\tilde{\gamma}\right)\right)'\left(\Sigma_i\Sigma_i'\right)\left(x^B\left(\tilde{\gamma},\beta\left(\tilde{\gamma}\right)\right)-\beta\left(\tilde{\gamma}\right)\right),\end{aligned} \tag{3.73}$$

that is, the relative risk of a more aggressive manager under a benchmark designed for a more conservative manager is larger than when the more conservative manager implements the strategy, since $\tilde{\gamma} > \gamma$. This implies that when the risk constraint is satisfied with equality for a manager with risk aversion $\tilde{\gamma}$, an unconstrained manager with risk aversion γ will implement a strategy that exceeds the relative risk limit. Consequently, the risk constraint on the basis of which the benchmark is designed will be binding and induces an effective increase in the manager's risk aversion from γ to $\tilde{\gamma}$.

3.D Tables and figures

Panel A: Mean-variance frontiers of the asset classes



Panel B: Centralized vs. decentralized mean-variance frontier

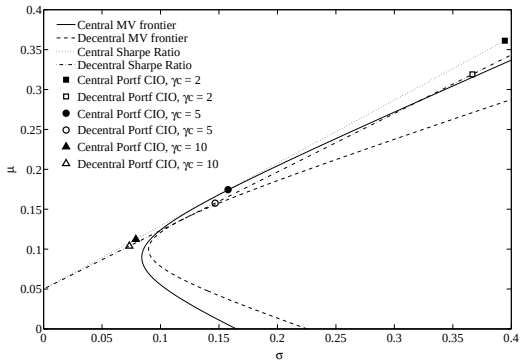
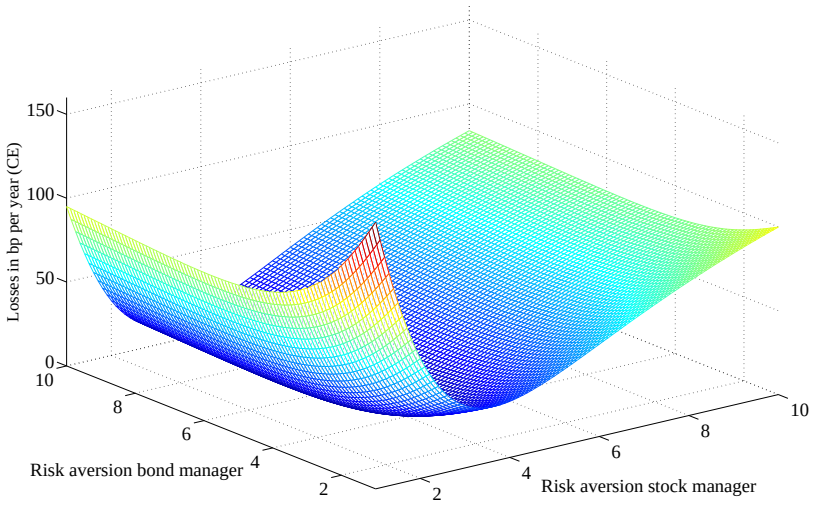


Figure 3.1: Decentralized investment management problem
This figure shows a decentralized asset allocation problem in which a CIO delegates portfolio decisions to a stock and a bond manager. Both asset managers have a risk aversion coefficient of $\gamma_1 = \gamma_2 = 10$. The bond manager invests in government bonds and corporate bonds with Aaa and Baa ratings. The stock manager invests in growth, intermediate, and value stocks. Panel A shows the mean-variance frontier for stocks and for bonds. The decentralized mean-variance frontier intersects the stock and bond mean-variance frontiers at the preferred portfolios of the bond and the stock manager. The CIO allocates money to the two managers and a riskless asset that pays 5% per year. Panel B compares the mean-variance frontier of the decentralized investment problem with that of the centralized investment problem and depicts the optimal portfolio choices of the CIO for the CIO's risk aversion level γ_C equal to 2, 5, and 10.

Panel A: The coefficient of relative risk aversion of the CIO equals $\gamma_C = 5$



Panel B: The coefficient of relative risk aversion of the CIO equals $\gamma_C = 10$

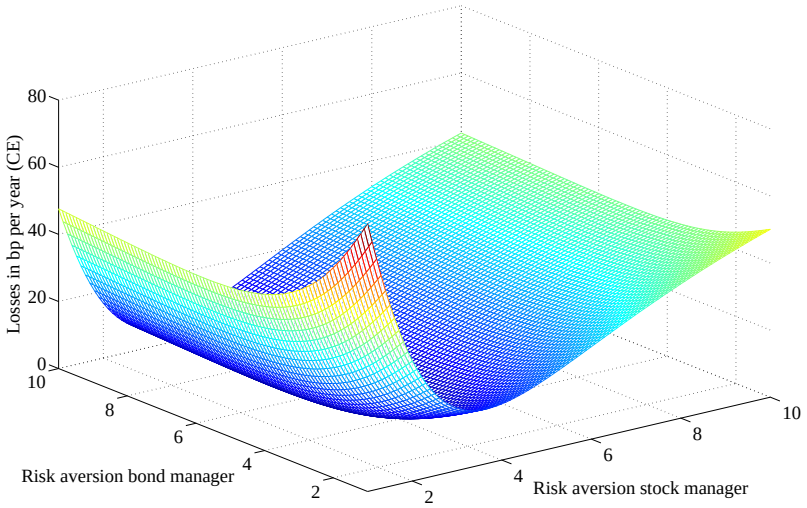
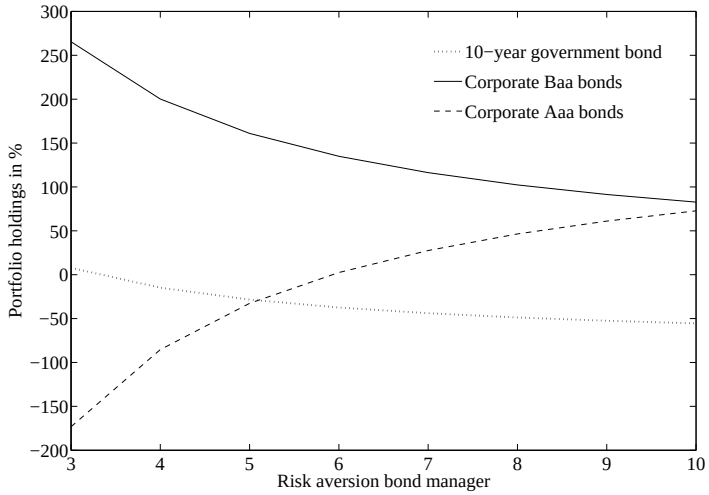


Figure 3.2: Losses from decentralized investment management
This figure depicts the diversification losses due to decentralized investment management as a function of the risk aversion of the investment managers. The CIO has a risk aversion coefficient $\gamma_C = 5$ in Panel A and $\gamma_C = 10$ in Panel B. The horizontal axes depict the risk appetites of the asset managers. The losses are computed by taking the ratio of the annualized certainty equivalents achieved under decentralized and centralized investment management after which we subtract one and multiply by -10,000 to express the losses in basis points per year. For example, 160 basis points implies a loss in terms of certainty equivalents of 1.6% of wealth per year.

Panel A: Portfolio composition bond manager



Panel B: Portfolio composition stock manager

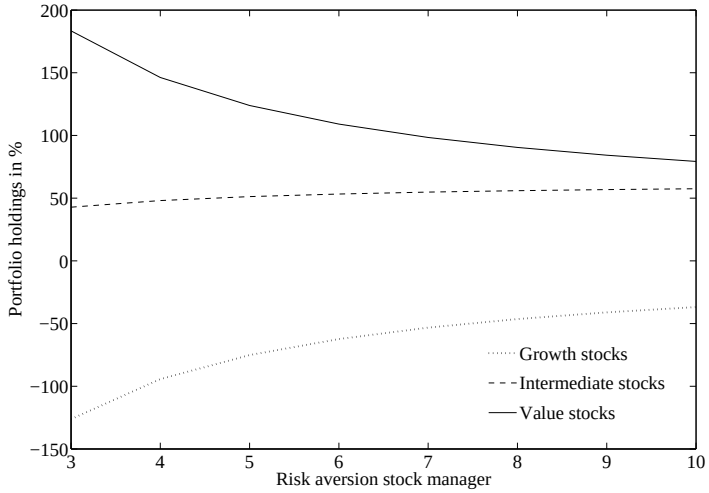


Figure 3.3: Portfolio compositions without a benchmark
This figure displays the portfolio composition of the bond manager in Panel A and the stock manager in Panel B as functions of their coefficients of relative risk aversion when they are not restricted by a benchmark. The asset managers do not have access to a riskless asset.

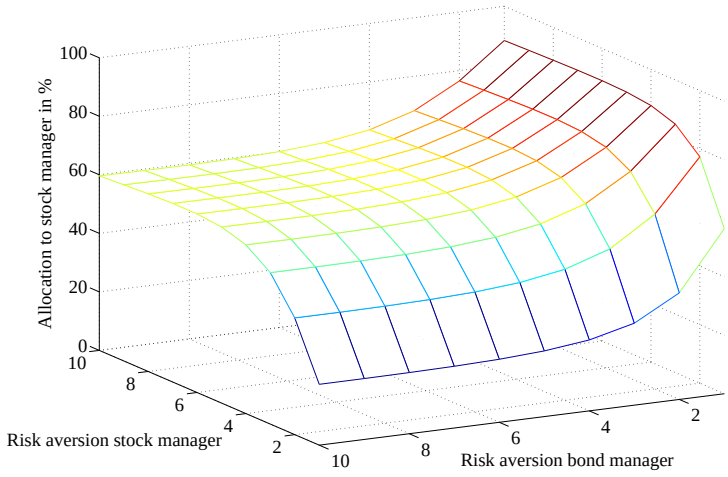
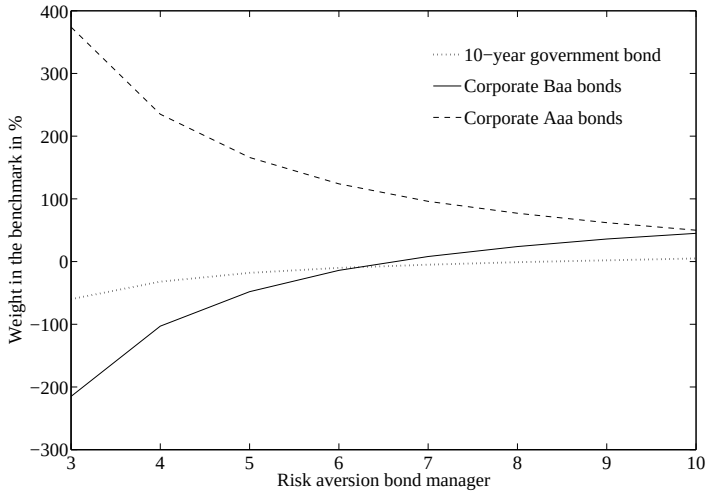


Figure 3.4: Fraction of risky funds allocated to equities without a benchmark
This figure displays the percentage of total investment in risky assets that is under control of the stock manager as a function of the risk aversion of the bond and the stock manager.

Panel A: Composition of the optimal bond benchmark



Panel B: Composition of the optimal stock benchmark

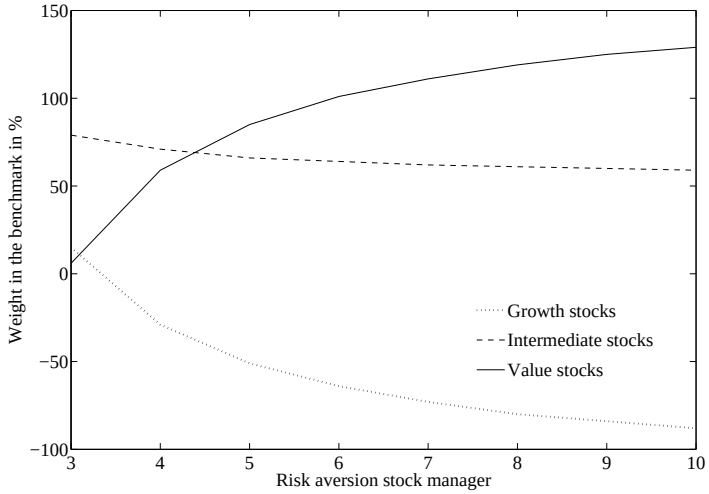
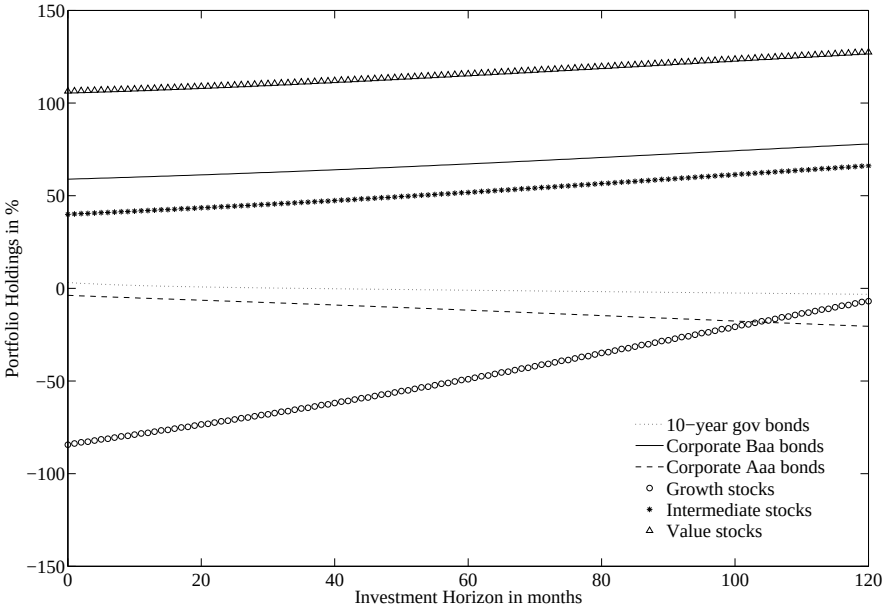


Figure 3.5: This figure gives the composition of the optimal performance benchmarks
Composition of the optimal bond benchmark in Panel A and stock benchmark in Panel B as a function of the risk aversion of the asset managers.

Panel A: The coefficient of relative risk aversion of the CIO equals $\gamma_C = 5$



Panel B: The coefficient of relative risk aversion of the CIO equals $\gamma_C = 10$

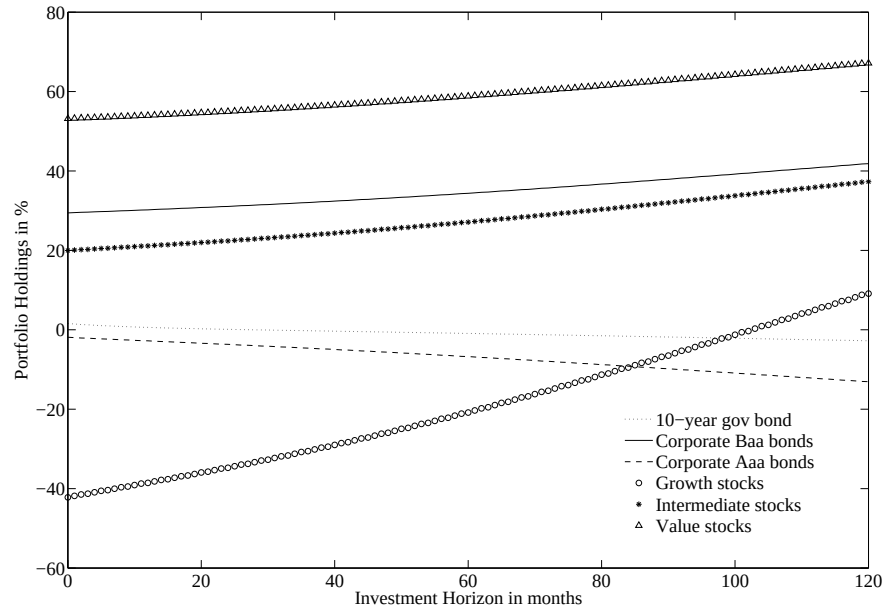


Figure 3.6: Optimal portfolio choice in the centralized problem
This figure depicts the optimal allocation to government bonds, corporate bonds with ratings Baa and Aaa, and three stock portfolios ranked based upon their book-to-market ratios (growth, intermediate, and value). The horizontal axis depicts the investment horizon of the CIO in months. The coefficient of relative risk aversion of the CIO equals $\gamma_C = 5$ in Panel A and $\gamma_C = 10$ in Panel B.

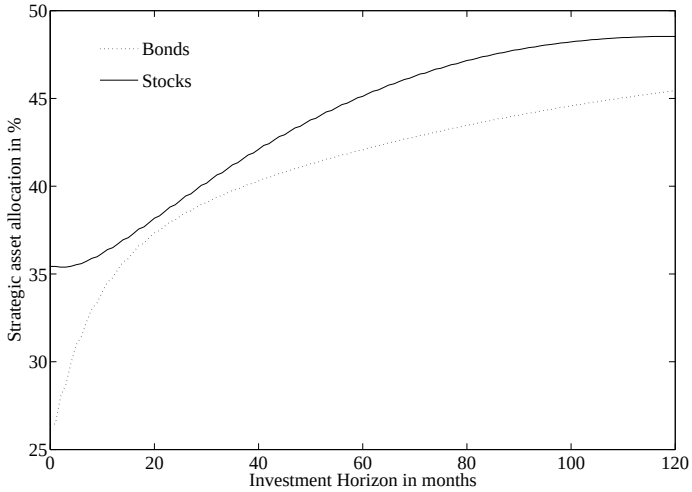
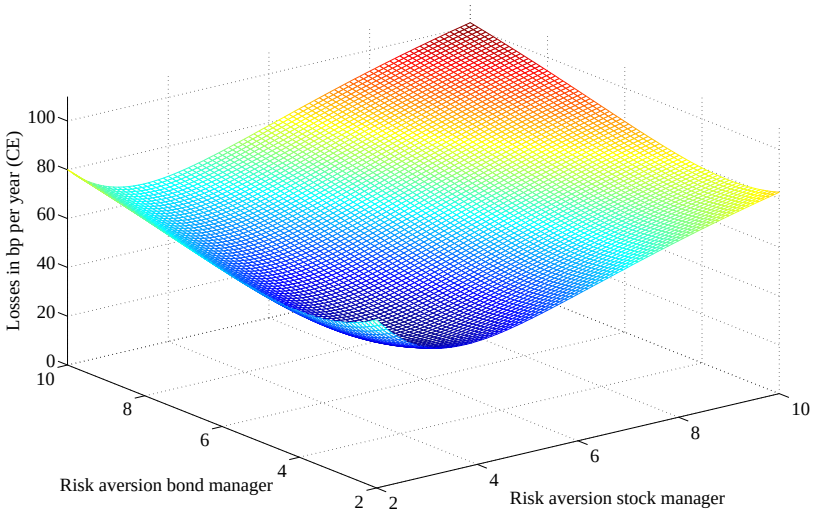


Figure 3.7: Optimal strategic allocation in the decentralized problem without a benchmark
This figure displays the optimal allocation to the fixed income and equity asset classes in absence of a benchmark. The horizontal axis depicts the investment horizon of the CIO in months. The preference parameters have been set to $\gamma_C = 10$ and $\gamma_i = 5$, with $i = 1, 2$.

Panel A: The investment horizon of the CIO equals $T = 1$



Panel B: The investment horizon of the CIO equals $T = 10$

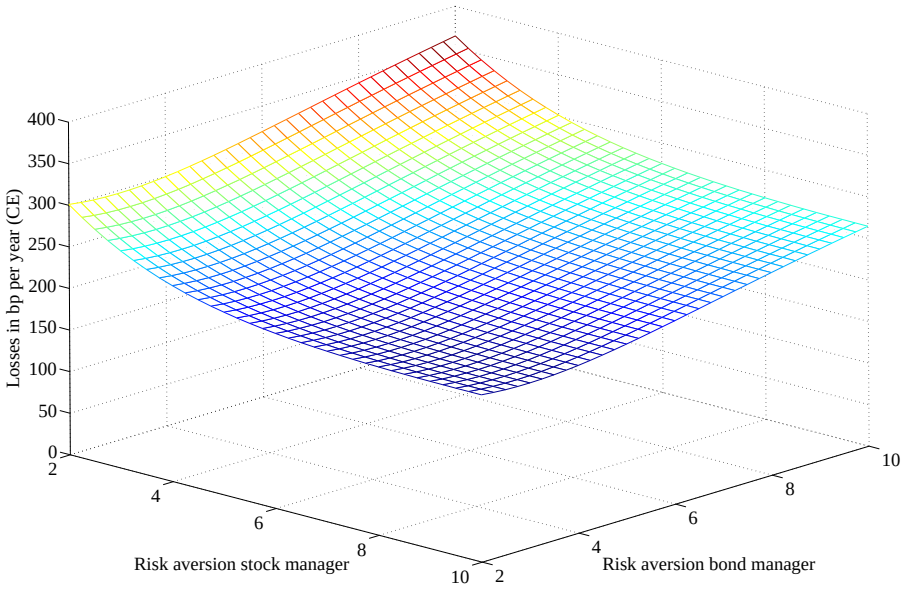
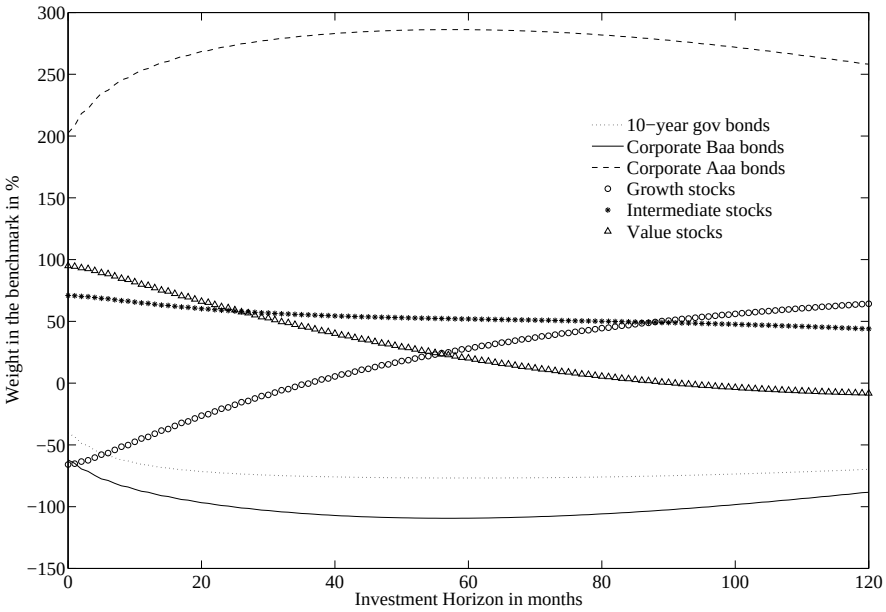


Figure 3.8: Utility costs of decentralized investment management without a benchmark
This figure gives a comparison of certainty equivalents following from the centralized and decentralized investment management problem when there is no benchmark and the investment horizon is one year in Panel A and 10 years in Panel B. The horizontal axes depict the risk appetites of the asset managers. The coefficient of relative risk aversion of the CIO equals 10. The losses are computed by taking the ratio of the annualized certainty equivalents achieved under decentralized and centralized investment management after which we subtract one and multiply by -10,000 to express the losses in basis points per year.

Panel A: Composition of the optimal performance benchmarks



Panel B: Optimal strategic asset allocation of the CIO

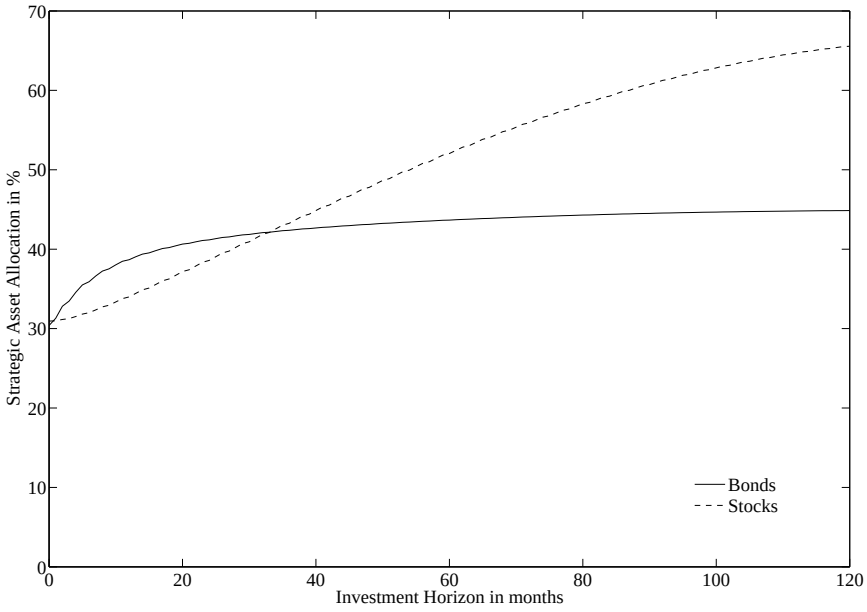
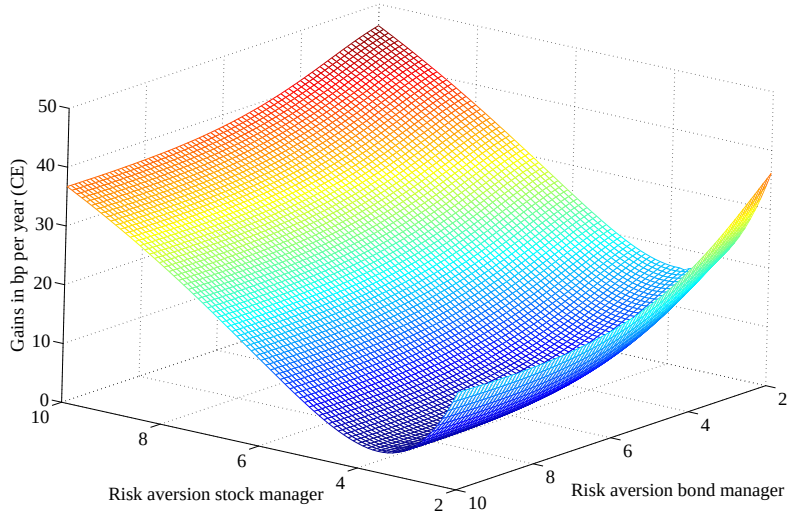


Figure 3.9: Optimal performance benchmarks and strategic allocation
Panel A portrays the composition of the optimal performance benchmarks for different investment horizons of the CIO. Panel B presents the corresponding optimal strategic asset allocation to the asset classes. We plot the benchmark for the stock and bond manager in the same graph, but there is still no cross-benchmarking. That is, the benchmark weights in both asset classes each sum up to 100%. The horizontal axis depicts the investment horizon of the CIO in months. The preference parameters are $\gamma_C = 10$ and $\gamma_i = 5$, with $i = 1, 2$.

Panel A: The investment horizon of the CIO equals $T = 1$



Panel B: The investment horizon of the CIO equals $T = 10$

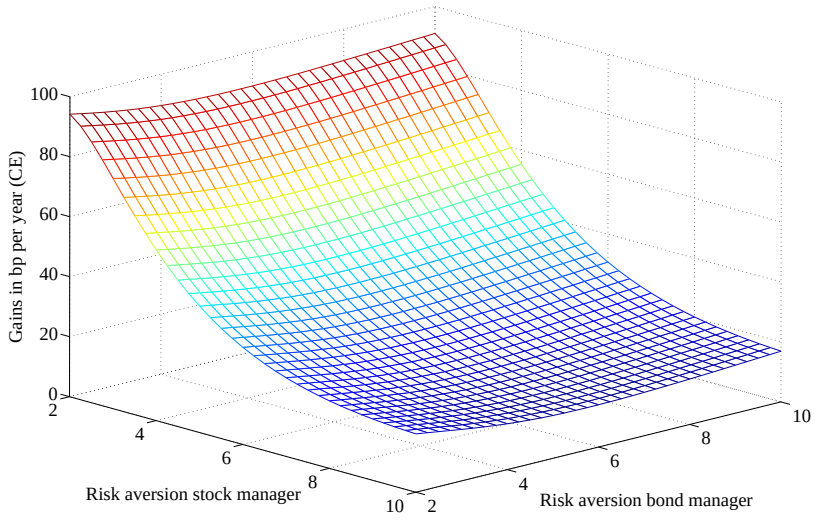


Figure 3.10: Value generated by an optimally chosen benchmark
This figure gives a comparison of certainty equivalents following from the decentralized with and without an optimally chosen benchmark. We present the annualized gains in basis points from using the benchmark optimally. The investment horizon of the CIO equals one year in Panel A and 10 years in Panel B. The horizontal axes depict different risk appetites of the asset managers. The coefficient of relative risk aversion of the CIO equals 10.

Panel A: Model parameters

Source of risk	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
Λ	0.331	0.419	-0.0291	0.126	0.477	0.305
Σ						
Gov. bonds	13.5%	0	0	0	0	0
Corp. bonds, Baa	8.2%	5.6%	0	0	0	0
Corp. bonds, Aaa	9.1%	2.7%	2.4%	0	0	0
Growth stocks	3.7%	6.3%	0.3%	16.5%	0	0
Int. stocks	3.6%	6.8%	0.3%	11.7%	7.3%	0
Value stocks	3.6%	7.7%	0.1%	10.4%	6.8%	5.9%

Panel B: Implied parameters

	Expected return	Correlation					
Gov. bonds	9.5%	100%	82%	93%	20%	23%	22%
Corp. bonds, Baa	10.1%	82%	100%	92%	37%	43%	45%
Corp. bonds, Aaa	9.1%	93%	92%	100%	29%	34%	34%
Growth stocks	10.9%	20%	37%	29%	100%	88%	80%
Int. stocks	14.0%	23%	43%	34%	88%	100%	93%
Value stocks	15.7%	22%	45%	34%	80%	93%	100%

Table 3.1: Constant investment opportunities
This table gives the estimation results of the financial market in Section 2 over the period January 1973 through November 2004 using monthly data. The model is estimated by maximum likelihood. The asset set contains government bonds ('Gov. bonds'), corporate bonds with credit ratings Baa ('Corp. bonds, Baa') and Aaa ('Corp. bonds, Aaa'), and three equity portfolio ranked on their book-to-market ratio (growth/intermediate ('Int. ')/value). Panel A provides the model parameters and Panel B portrays the implied instantaneous expected returns ($r + \Sigma\Lambda$) and correlations. In determining Λ, we assume that the instantaneous nominal short rate equals $r = 5\%$.

Source of risk	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
Λ_0	0.306	0.409	-0.020	0.089	0.498	0.310	0	0	0

	$\Sigma\Lambda_1$						
	Gov.	Baa	Aaa	Growth	Int.	Value	κ_i
Short rate	0.227	-0.964	-0.209	-0.270	-0.249	-0.012	0.36
10Y yield	1.269	1.225	0.893	-0.778	-1.086	-1.010	0.12
DP	0.020	0.071	0.038	0.132	0.121	0.130	0.052

Σ	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9
Gov. bonds	13.2%	0	0	0	0	0	0	0	0
Corp. bonds, Baa	7.7%	5.4%	0	0	0	0	0	0	0
Corp. bonds, Aaa	8.7%	2.6%	2.4%	0	0	0	0	0	0
Growth stocks	3.1%	5.8%	0.2%	16.5%	0	0	0	0	0
Int. stocks	2.9%	6.2%	0.1%	11.7%	7.2%	0	0	0	0
Value stocks	2.8%	7.1%	-0.2%	10.4%	6.7%	5.8%	0	0	0
Short rate	-1.1%	-0.1%	0.0%	0.3%	-0.1%	-0.1%	2.3%	0	0
10Y yield	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%	1.3%	0
DP	-3.0%	-6.7%	0.1%	-14.0%	-2.5%	-0.9%	0.0%	0.6%	4.7%

Table 3.2: Time-varying investment opportunities
This table shows the estimation results of the financial market in Section 3 over the period January 1973 through November 2004 using monthly data. The model is estimated by maximum likelihood. The asset set contains government bonds ('Gov. bonds'), corporate bonds with credit ratings Baa ('Corp. bonds, Baa') and Aaa ('Corp. bonds, Aaa'), and three equity portfolio ranked on their book-to-market ratio (growth/intermediate ('Int. ')/value). In determining Λ_0, we assume that the instantaneous nominal short rate equals $r = 5\%$. We report $\Sigma\Lambda_1$ rather than Λ_1 as the former expression is easier to interpret. The short rate, the yield on a 10Y nominal government bond, and the dividend yield are used to predict returns.

Panel A: Constant investment opportunities

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	19%	30%	22% (22%)	36% (36%)	24%	37%
$\sigma_\gamma = 1$	18%	27%	22% (22%)	36% (36%)	23%	37%
$\sigma_\gamma = 2$	18%	26%	21% (21%)	33% (33%)	23%	36%
$\sigma_\gamma = 3$	18%	27%	21% (21%)	31% (31%)	22%	33%
$\sigma_\gamma = 25$ (uniform)	20%	28%	20% (20%)	28% (29%)	20%	28%

Panel B: Time-varying investment opportunities ($T = 1$)

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	27%	31%	36%	37%	39%	38%
$\sigma_\gamma = 1$	23%	27%	35%	37%	38%	38%
$\sigma_\gamma = 2$	22%	26%	29%	34%	35%	37%
$\sigma_\gamma = 3$	22%	27%	27%	31%	30%	35%
$\sigma_\gamma = 25$ (uniform)	24%	29%	24%	29%	24%	29%

Panel C: Time-varying investment opportunities ($T = 10$)

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	33%	35%	47%	51%	51%	57%
$\sigma_\gamma = 1$	23%	26%	26%	42%	51%	56%
$\sigma_\gamma = 2$	23%	25%	25%	30%	26%	36%
$\sigma_\gamma = 3$	23%	25%	25%	28%	25%	31%
$\sigma_\gamma = 25$ (uniform)	24%	26%	24%	26%	24%	26%

Table 3.3: Strategic allocation without benchmarks when risk attitudes are unknown
This table gives the strategic allocation of the CIO to the asset classes when the risk attitudes of the managers are unknown and there are no benchmarks. The prior of the CIO over the risk aversion level of each of the managers is a truncated normal distribution with parameters μ_γ and σ_γ, truncated below at one and truncated above at 10.

Panel A: Constant investment opportunities

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	13.8	6.9	29.8	14.9	54.6	27.2
$\sigma_\gamma = 1$	51.5	25.7	31.7	15.9	53.8	26.9
$\sigma_\gamma = 2$	73.8	36.9	51.4	25.7	53.0	26.5
$\sigma_\gamma = 3$	80.5	40.2	68.3	34.2	62.2	31.1
$\sigma_\gamma = 25$ (uniform)	86.8	43.4	86.8	43.4	86.8	43.4

Panel B: Time-varying investment opportunities ($T = 1$)

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	94.4	47.4	119.7	58.9	158.6	78.0
$\sigma_\gamma = 1$	152.5	76.7	124.1	61.2	157.9	77.7
$\sigma_\gamma = 2$	188.6	94.7	161.4	80.4	159.8	78.9
$\sigma_\gamma = 3$	201.9	101.2	189.6	94.7	179.8	89.4
$\sigma_\gamma = 25$ (uniform)	217.6	108.8	217.6	108.8	217.6	108.8

Panel C: Time-varying investment opportunities ($T = 10$)

	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	434.0	261.1	401.0	234.3	434.6	245.1
$\sigma_\gamma = 1$	586.2	341.2	503.0	295.6	439.3	248.2
$\sigma_\gamma = 2$	650.7	372.6	633.0	363.5	611.5	351.8
$\sigma_\gamma = 3$	679.4	386.4	679.0	386.0	669.9	381.3
$\sigma_\gamma = 25$ (uniform)	717.4	404.6	717.4	404.6	717.4	404.6

Table 3.4: Costs of decentralized investment management if risk attitudes are unknown
This table gives the costs of decentralized investment management when the risk attitudes of the managers are unknown and there are no benchmarks. The prior of the CIO over the risk aversion levels of each of the managers is a truncated normal distribution with parameters μ_γ and σ_γ, truncated below at one and truncated above at 10. The losses are computed by taking the ratio of the annualized certainty equivalents achieved under decentralized and centralized investment management after which we subtract one and multiply by -10,000 to express the losses in basis points per year.

	Panel A: Constant investment opportunities					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	24%	32%	24%	32%	24%	32%
$\sigma_\gamma = 1$	21%	26%	24%	31%	24%	32%
$\sigma_\gamma = 2$	19%	24%	23%	29%	24%	31%
$\sigma_\gamma = 3$	19%	24%	21%	27%	23%	29%
$\sigma_\gamma = 25$ (uniform)	20%	25%	20%	25%	20%	25%
	Panel B: Time-varying investment opportunities ($T = 1$)					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	31%	35%	40%	34%	41%	33%
$\sigma_\gamma = 1$	25%	28%	38%	33%	41%	33%
$\sigma_\gamma = 2$	23%	25%	31%	30%	37%	33%
$\sigma_\gamma = 3$	23%	25%	27%	28%	31%	31%
$\sigma_\gamma = 25$ (uniform)	24%	26%	24%	26%	24%	26%
	Panel C: Time-varying investment opportunities ($T = 10$)					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	Bonds	Stocks	Bonds	Stocks	Bonds	Stocks
$\sigma_\gamma = 0$	34%	61%	47%	66%	50%	67%
$\sigma_\gamma = 1$	23%	28%	26%	43%	50%	66%
$\sigma_\gamma = 2$	23%	25%	25%	30%	26%	36%
$\sigma_\gamma = 3$	23%	25%	25%	27%	25%	30%
$\sigma_\gamma = 25$ (uniform)	24%	25%	24%	25%	24%	25%

Table 3.5: Strategic allocation with benchmarks when risk attitudes are unknown
This table gives the strategic allocation of the CIO to the asset classes when the risk attitudes of the managers are unknown and the optimal benchmarks are implemented. The prior of the CIO over the risk aversion levels of each of the managers is a truncated normal distribution with parameters μ_γ and σ_γ, truncated below at one and truncated above at 10.

	Panel A: Constant investment opportunities					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	13.8	6.9	29.8	14.9	54.6	27.2
$\sigma_\gamma = 1$	9.1	4.5	28.5	14.3	53.1	26.5
$\sigma_\gamma = 2$	13.4	6.7	29.3	14.6	46.7	23.3
$\sigma_\gamma = 3$	19.8	9.9	31.2	15.6	41.3	20.6
$\sigma_\gamma = 25$ (uniform)	33.7	16.9	33.7	16.9	33.7	16.9
	Panel B: Time-varying investment opportunities ($T = 1$)					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	28.2	14.5	29.4	13.9	48.6	23.2
$\sigma_\gamma = 1$	12.3	6.0	28.0	13.3	47.3	22.6
$\sigma_\gamma = 2$	13.5	6.5	26.5	12.7	41.6	19.8
$\sigma_\gamma = 3$	18.7	9.0	28.2	13.6	36.7	17.6
$\sigma_\gamma = 25$ (uniform)	31.2	15.1	31.2	15.1	31.2	15.1
	Panel C: Time-varying investment opportunities ($T = 10$)					
	$\mu_\gamma = 3.1$		$\mu_\gamma = 5.5$		$\mu_\gamma = 7.3$	
	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$	$\gamma_C = 5$	$\gamma_C = 10$
$\sigma_\gamma = 0$	110.3	71.5	40.4	31.5	26.7	20.4
$\sigma_\gamma = 1$	14.3	7.6	21.4	11.2	26.8	20.4
$\sigma_\gamma = 2$	11.5	6.0	18.2	9.3	23.7	12.2
$\sigma_\gamma = 3$	14.7	7.4	20.3	10.2	24.6	12.3
$\sigma_\gamma = 25$ (uniform)	22.8	11.3	22.8	11.3	22.8	11.3

Table 3.6: Value of optimal benchmarks when risk attitudes are unknown
This table gives a comparison of certainty equivalents following from the decentralized with and without an optimally chosen benchmark. We present the annualized gains in basis points from using the benchmark optimally. The prior of the CIO over the risk aversion levels of each of the managers is a truncated normal distribution with parameters μ_γ and σ_γ, truncated below at one and truncated above at 10.

Chapter 4

Mortgage Timing

Abstract

We study how the term structure of interest rates relates to mortgage choice, both at the household and the aggregate level. A simple utility framework of mortgage choice points to the long-term bond risk premium as theoretical determinant: when the bond risk premium is high, fixed-rate mortgage payments are high, making adjustable-rate mortgages more attractive. This long-term bond risk premium is markedly different from other term structure variables that have been proposed, including the yield spread and the long yield. We confirm empirically that the bulk of the time variation in both aggregate and loan-level mortgage choice can be explained by time variation in the bond risk premium. This is true whether bond risk premia are measured using forecasters' data, a VAR term structure model, or from a simple household decision rule based on adaptive expectations. This simple rule moves in lock-step with mortgage choice, lending credibility to a theory of strategic mortgage timing by households.

4.1 Introduction

One of the most important financial decisions any household has to make during its lifetime is whether to own a house and, if so, how to finance it. There are two broad categories of housing finance: adjustable-rate mortgages (ARMs) and fixed-rate mortgages (FRMs). The share of newly-originated mortgages that is of the ARM-type in the US economy shows a surprisingly large variation. It varies between 10% and 70% of all mortgages over our sample period from January 1985 to June 2006. We seek to understand these fluctuations in the ARM share.

The main contribution of our paper is to understand the link from the term structure of interest rates to both individual and aggregate mortgage choice. While various term structure variables, such as the yield spread and the long-term yield (e.g., Campbell and Cocco (2003)), have been proposed before, the literature lacks a theory that predicts the precise link between the term structure and mortgage choice. A simple utility framework allows us to show that the long-term bond risk premium is the key determinant. This is the

premium earned on investing long in a long-term bond and rolling over a short position in short-term bonds. The premium arises whenever the expectations hypothesis of the term structure of interest rates fails to hold, a fact for which there is abundant empirical evidence by now. We are the first to propose the bond risk premium as a predictor of mortgage choice and to document its strong predictive ability. We show that the long-term bond risk premium is conceptually and empirically very different from both the yield spread and the long yield. Because both variables are imperfect proxies for the long-term bond risk premium, they are imperfect predictors of mortgage choice.

What makes the bond risk premium a palatable determinant of observed household mortgage choice? Imagine a household which has to choose between an FRM and an ARM to finance its house purchase. With an FRM, mortgage payments are constant and linked to the long-term interest rate at the time of origination. With an ARM, matters are more complicated: future ARM payments will depend on future short-term interest rates not known at origination. We imagine that the household uses an average of short-term interest rates from the recent past in order to estimate future ARM payments. Under such expectations-formation rule, the difference between the long-term interest rate and the recent average of short-term interest rates is what the household would use to make the choice between the FRM and the ARM. Therefore, we label this difference the *household's decision rule*. The theoretical long-term bond risk premium that follows from our model is the -closely related- difference between the current long yield and the average expected future short yields over the contract period. The household decision rule is a proxy for the bond risk premium which arises when adaptive expectations are formed. Our motivation for this approximation is a suspicion that households may not have the required financial sophistication to solve complex investment problems (Campbell (2006)). The household decision rule is easy to compute, conceptually intuitive, and theoretically-founded.

This simple rule is highly effective at choosing the right mortgage at the right time. Section 4.2 shows that it has a correlation of 81% with the observed ARM share in the aggregate time series. We also use a new, nation-wide, loan-level data set that allows us to link the household decision rule to several hundred thousand individual mortgage choices. We find that it alone classifies 70% of mortgage loans correctly. The marginal impact of the household decision rule is essentially unaffected once we control for loan-level characteristics and geographic variables. In fact, the rule is an economically more significant predictor of individual mortgage choice than various individual-specific measures of financial constraints. The loan-level data reiterate the problem with the yield spread and the long yield as predictors of mortgage choice.

Section 4.3 presents our model; its novel feature is allowing for time variation in bond

risk premia. The model is kept deliberately simple, as in Campbell (2006), and strips out some of the rich life-cycle dynamics modeled elsewhere.[1] It models risk averse households who trade off the expected payments on an FRM and an ARM contract with the risk of these payments. The ARM payments are subject to real interest rate risk, while the presence of inflation uncertainty makes the real FRM payments risky. The model generates an intuitive risk-return trade-off for mortgage choice: the ARM contract is more desirable the higher the nominal bond risk premium, the lower the variability of the real rate, and the higher the variability of expected inflation. We explicitly aggregate the mortgage choice across households that are heterogeneous in risk preferences. Time variation in the aggregate ARM share is then caused by time variation in the bond risk premium. The mean and dispersion parameters of the cross-sectional distribution of risk aversion map one-to-one into the average ARM share and its sensitivity to the bond risk premium, respectively. The model also helps us understand the problem with the yield spread and long yield as predictors of mortgage choice. The yield spread is a noisy proxy for the long-term bond risk premium because average expected future short rates differ from the current short rate due to mean reversion. This creates an errors-in-variables problem in the regression of the ARM share on the yield spread. The problem is so severe in the data that the yield spread is effectively uninformative about the future ARM share. Intuitively, the yield spread fails to take into account that future ARM payments will adjust whenever the short rate changes. A similar, though empirically less pronounced, errors-in-variables problem occurs for the long yield.

In Section 4.4, we bring the theory to the data, and regress the ARM share on the nominal bond risk premium. We first show formally that the household decision rule arises as a measure of the bond risk premium when expectations of future nominal short rates are computed with an adaptive expectations scheme. This provides the theoretical underpinning for the empirical success of the household decision rule in predicting mortgage choice. The simple proxy for the bond risk premium explains about 70% of the variation in the ARM share. We also explore more academically conventional ways of measuring expected future short rates: based on Blue Chip forecasters' data and based on a vector auto-regression model of the term structure. These two forward-looking bond risk premia measures generate the same quantitative sensitivity of the ARM share: a one standard deviation increase in the bond risk premium leads to an 8% increase in the ARM share. This is a large economic effect given the average ARM share of 28%.

While the forward-looking measures of the bond risk premium deliver similar results to the household decision rule over the full sample, their performance diverges in the last ten years of the sample. This is mostly due to the increase in the ARM share in 2003-04, which

[1]For instance, Campbell and Cocco (2003), Cocco (2005), and Van Hemert (2006).

is predicted correctly by the simple rule, but not by the other two forward-looking measures of the bond risk premium. Section 4.5 explains this divergence. Part of the explanation lies in product innovation in the ARM mortgage segment. But most of the divergence is due to large forecast errors in future short rates in this episode. This motivates us to consider the *inflation* risk premium component of the nominal risk premium, for which any forecast error that is common to nominal and real rates cancels out. We construct the inflation risk premium using real yield (TIPS) data and either Blue Chip forecasters' data or a VAR model for inflation expectations, and show that both measures have a strong positive correlation with the ARM share and deliver a similar economic effect.

In Section 4.6, we extend our baseline results. First, we analyze the impact of the prepayment option, typically embedded in US FRM contracts, on the utility difference between the ARM and FRM. We show that the prepayment option reduces the exposures to the underlying risk factors. However, it continues to hold that higher bond risk premia favor ARMs. In sum, we find that the presence of the option does not materially alter the results. Second, we investigate the role of financial constraints using aggregate and loan-level data. The loan level data allow us to investigate the importance of measures of financial constraints, such as the loan-to-value ratio or the credit score, for the relative desirability of the ARM. While they are statistically significant predictors of mortgage choice, they do not add much to the explanatory power of the bond risk premium, nor significantly reduce it. In the context of financial constraints, we also investigate the role of short investment horizons as captured by a high rate of impatience or a high moving probability in a dynamic version of our model. When households are so impatient or have such high moving probability that they only care about the first mortgage payment, the yield spread fully captures the FRM-ARM tradeoff. For realistic values for moving rates or rates of time preference, the bond risk premium is the relevant determinant. Fourth, we discuss the robustness of the statistical inference, and conduct a bootstrap exercise to calculate standard errors. Finally, we discuss liquidity issues in the TIPS markets and how they may affect our results on the inflation risk premium. We conclude that bond risk premia are a robust determinant of mortgage choice.

Our findings resonate with recent work in the portfolio literature by Campbell, Chan, and Viceira (2003), Sangvinatsos and Wachter (2005), Brandt and Santa-Clara (2006), and Koijen, Nijman, and Werker (2007a). This literature emphasizes that forming portfolios that take into account time-varying risk premia can substantially improve performance for long-term investors.[2] Because the mortgage is a key component of the typical household's portfolio, and because an ARM exposes that portfolio to different interest rate risk than an

[2]Campbell and Viceira (2001b) and Brennan and Xia (2002) derive the optimal portfolio strategy for long-term investors in the presence of stochastic real interest rates and inflation, but assume risk premia to be constant.

FRM, choosing the wrong mortgage may have adverse welfare consequences (Campbell and Cocco (2003) and Van Hemert (2006)). In contrast to these studies, our exercise suggests that mortgage choice is an important financial decision where the use of bond risk premia is not only valuable from a normative point of view. Time variation in risk premia is also important from a positive point of view, to explain observed variation in mortgage choice both at the aggregate and at the household level.

Finally, our paper also relates to the corporate finance literature on the timing of capital structure decisions. The firm's problem of maturity choice of debt is akin to the household's choice between an ARM and an FRM. Baker, Greenwood, and Wurgler (2003) show that firms are able to time bond markets. The maturity of debt decreases in periods of high bond risk premia.[3] Our findings suggest that households also have the ability to incorporate information on bond risk premia in their long-term financing decision.

4.2 A Simple Story for Household Mortgage Choice

We imagine a household that is choosing between a standard fixed-rate and a standard adjustable-rate mortgage contract. On the FRM contract, it will pay a fixed, long-term interest rate while the rate on the ARM contract will reset periodically depending on the short-term interest rate. The household knows the current long-term interest rate, but lacks a sophisticated model for predicting future short-term interest rates. Instead, it naively forms an average of the short rate over the recent past as a proxy of what it expects to pay on the ARM. The relative attractiveness of the ARM contract is the difference between the current long rate and the average short rate over the recent past. We label this difference at time t the *household decision rule* κ_t.

Figure 4.1 displays the time series of the share of newly-originated mortgages that is of the ARM type (solid line, left axis) alongside the household decision rule $\kappa_t(3,5)$ (dashed line, right axis). The latter is formed using the 5-year Treasury bond yield (indicated by the second argument) and the 1-year Treasury bill yield averaged over the past three years (indicated by the first argument). The ARM share is from the Federal Housing Financing Board, the standard source in the literature. Appendix 4.A discusses the data in more detail and compares it to other available series. The figure documents a striking co-movement between the ARM share and the decision rule; their correlation is 81%. In Section 4.4 below, we present similar evidence from a regression analysis.

[3]See Butler, Grullon, and Weston (2006) and Baker, Taliaferro, and Wurgler (2006) for a recent discussion. In ongoing work, Greenwood and Vayanos (2007) study the the relationship between government bond supply and excess bond returns.

Figure 4.2 shows that this high correlation not only holds when the household decision rule is formed using Treasury interest rates (left panel), but also using mortgage interest rates (right panel). In both panels the household decision rule κ has the strongest association with the ARM share (highest bar) for intermediate values of the horizon over which average short rates are computed. The correlation is hump-shaped in the look-back horizon.

We not only find such high correlation between the household decision rule and the ARM share in aggregate time series data, but also in individual loan-level data. We explore a new data set which contains information on 911,000 loans from a large mortgage trustee for mortgage-backed security special purpose vehicles. The loans were issued between 1994 and 2007.[4] Table 4.1 reports loan-level results of probit regressions with an ARM dummy as left-hand side variable. All right-hand side variables have been scaled by their standard deviation. We report the coefficient estimate, a robust t-statistic, and the fraction of loans that is correctly classified by the probit model.[5] We keep the 654,368 loans for which we have all variables of interest available. The first row shows that the household decision rule is a strong predictor of loan-level mortgage choice. It has the right sign, a t-statistic of 253, and it -alone- classifies 69.4% of loans correctly. Its coefficient indicates that a one standard deviation increase in the bond risk premium increases the probability of an ARM choice from 39% to 56%, an increase of more than one-third.

It is interesting to contrast this result with a similar probit regression that has three well-documented indicators of financial constraints on the right-hand side: the loan balance at origination (BAL), the credit score of the borrower (FICO), and the loan-to-value ratio (LTV). The second row, which also includes four regional dummies for the biggest mortgage markets (California, Florida, New York, and Texas), confirms that a lower balance, a lower FICO score, and especially a higher LTV ratio increase the probability of choosing an ARM. However, the (scaled) coefficients on the loan characteristics are smaller than the coefficient on the household decision rule κ, suggesting a smaller economic effect. Furthermore, the three financial constraint variables classify only 59.0% of loans correctly; adding four state dummies increases correct classifications to 61.7%. Adding the three financial constraint proxies and the four regional dummies to the household decision rule does not increase the probability of classified loans (Row 3). The number of classified loans is 68.8%, no bigger than what is explained by κ alone.[6] Moreover, the household decision rule variable remains the largest and by far the most significant regressor. Its marginal effect on the probability of choosing an ARM is unaffected.

[4] Appendix 4.A provides more detail. We thank Nancy Wallace for graciously making these data available to us.

[5] By pure chance, one would classify 50% of the contracts correctly.

[6] Note that the maximum likelihood estimation does not maximize correct classifications, so that adding regressors does not necessarily increase correct classifications.

The rest of the paper is devoted to understanding why the simple decision rule works. We argue that it is a good proxy for the bond risk premium. The next section develops a rational model of mortgage choice that links time variation in the bond risk premium to time variation in the ARM share. While households might not have the required financial sophistication to solve complex investment problems (Campbell (2006)), the near-optimality of the simple decision rule suggests that close-to-rational mortgage decision making may well be within reach.[7]

The bond risk premium is not to be confused with the yield spread, which is the difference between the current long yield and the current short yield. To illustrate this distinction, the household decision rule in Figure 4.1 has a correlation of -25% with the 5-1 year yield spread. While κ had a correlation with the ARM share of 81%, the correlation between the yield spread and the aggregate ARM share is -6% over the same sample. This correlation is indicated by the solid line in the left panel of Figure 4.2. The correlation with the mortgage rate spread, indicated by the solid line in the right panel, is somewhat higher at 33%. However, it remains substantially below the 81% of the simple rule with mortgage rates. The long yield also has a much lower correlation with the ARM share than the household decision rule (dashed lines). The second role of the model is to help clarify the distinction between the bond risk premium and the yield spread or long yield.

4.3 Model with Time-Varying Bond Risk Premia

Various term structure variables have been suggested in the literature to predict aggregate mortgage choice, such as the yield spread and yields of various maturities.[8] The question of which term structure variable is the best predictor of individual and aggregate mortgage choice motivates us to set up a model that explores this link. Rather than developing a full-fledged life-cycle model, we study a tractable two-period model that allows us to focus solely on the role of time variation in bond risk premia. This extension of Campbell (2006) is motivated by the empirical evidence pointing to the failure of the expectations hypothesis in US post-war data.[9] We first explore an individual household's choice between a fixed-rate mortgage (FRM) and an adjustable-rate mortgage (ARM) (Sections 4.3.1-4.3.4). Subsequently, we aggregate mortgage choices across households to link the term structure

[7]One branch of the real estate finance literature documents slow prepayment behavior (e.g., Schwartz and Torous (1989)). Brunnermeier and Julliard (2006) study the effect of money illusion on house prices, and Gabaix, Krishnamurthy, and Vigneron (2006) study limits to arbitrage in mortgage-backed securities markets.

[8]For instance, Berkovec, Kogut, and Nothaft (2001), Campbell and Cocco (2003), and Vickery (2006).

[9]Fama and French (1989), Campbell and Shiller (1991), Dai and Singleton (2002), Buraschi and Jiltsov (2005), Ang and Piazzesi (2003), Cochrane and Piazzesi (2005), and Ang, Bekaert, and Wei (2006), among others, document and study time variation in bond risk premia.

dynamics to the ARM share (Section 4.3.6). The model sheds light on the difference between the bond risk premium, the yield spread, and the long yield in Section 4.3.5. Finally, Section 4.3.7 discusses extensions of the model and the relationship with the literature.

4.3.1 Setup

We consider a continuum of households on the unit interval, indexed by j. Households are identical, except in their attitudes toward risk parameterized by γ_j. The cumulative distribution function of risk aversion coefficients is denoted by $F(\gamma)$.

At time 0, households purchase a house and use a mortgage to finance it. The house has a nominal value $H_t^{\$}$ at time t. For simplicity, the loan is non-amortizing. We assume a loan-to-value ratio equal to 100%, so that the mortgage balance is given by $B = H_0^{\$}$. The investment horizon and the maturity of the mortgage contract equal 2 periods. Interest payments on the mortgage are made at times 1 and 2. At time $t = 2$, the household sells the house at a price $H_2^{\$}$ and pays down the mortgage. The household chooses to finance the house using either an ARM or an FRM, with associated nominal interest rates q^i, $i \in \{ARM, FRM\}$. In each period, the household receives nominal income $L_t^{\$}$.

We postulate that the household is borrowing constrained: In each period, she consumes what is left over from the income she receives after making the mortgage payment (equation (4.2)). Because the constrained household cannot invest in the bond market, she cannot undo the position taken in the mortgage market. Terminal consumption equals income after the mortgage payment plus the difference between the value of the house and the mortgage balance (equation (4.3)).

Each household maximizes lifetime utility over real consumption streams $\{C/\Pi\}$, where Π is the price index and $\Pi_0 = 1$. Preferences in (4.1) are of the CARA type with risk aversion parameter γ_j, except for a log transformation. The subjective time discount factor is $\exp(-\beta)$.[10] The maximization problem of household j reads:

$$\max_{i \in \{ARM,FRM\}} - \log\left(\mathbb{E}_0\left[e^{-\beta-\gamma_j \frac{C_1}{\Pi_1}}\right]\right) - \log\left(\mathbb{E}_0\left[e^{-2\beta-\gamma_j \frac{C_2}{\Pi_2}}\right]\right) \tag{4.1}$$

$$\text{s.t.} \quad C_1 = L_1^{\$} - q_1^i B, \tag{4.2}$$

$$C_2 = L_2^{\$} - q_2^i B + H_2^{\$} - B. \tag{4.3}$$

[10] This log transformation is reminiscent of an Epstein and Zin (1989) aggregator which introduces a small preference for early resolution of uncertainty (see also Van Nieuwerburgh and Veldkamp (2006)). While this modification is solely made for analytical convenience, it implies that β does not affect mortgage choice. In Section 4.6.2, we investigate the role of the subjective discount rate in a calibrated, multi-period model with CRRA preferences. We show that the risk-return tradeoff which governs mortgage choice is unaffected for conventional values of β. The same conclusion holds when we introduce a realistic moving rate.

We assume that real labor income, $L_t = L_t^{\$}/\Pi_t$, is stochastic and persistent:

$$L_{t+1} = \mu_L + \rho_L (L_t - \mu_L) + \sigma_L \varepsilon_{t+1}^L, \varepsilon_{t+1}^L \sim \mathcal{N}(0,1).$$

In addition, we assume that the real house value is constant and let $H_t = H_t^{\$}/\Pi_t$.

4.3.2 Bond Pricing

The one-period nominal short rate at time t, $y_t^{\$}(1)$, is the sum of the real rate, $y_t(1)$, and expected inflation, x_t:

$$y_t^{\$}(1) = y_t(1) + x_t. \tag{4.4}$$

Denote the corresponding price of the one-period nominal bond by $P_t^{\$}(1)$. Following Campbell and Cocco (2003), we assume that realized inflation and expected inflation coincide:

$$\pi_{t+1} = \log \Pi_{t+1} - \log \Pi_t = x_t, \tag{4.5}$$

so that there is no unexpected inflation risk.[11] To accommodate the persistence in the real rate and expected inflation, we model both processes to be first-order autoregressive:

$$\begin{aligned} y_{t+1}(1) &= \mu_y + \rho_y (y_t(1) - \mu_y) + \sigma_y \varepsilon_{t+1}^y, \\ x_{t+1} &= \mu_x + \rho_x (x_t - \mu_x) + \sigma_x \varepsilon_{t+1}^x. \end{aligned}$$

Their innovations are jointly Gaussian with correlation matrix R:

$$\begin{pmatrix} \varepsilon_{t+1}^y \\ \varepsilon_{t+1}^x \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{xy} \\ \rho_{xy} & 1 \end{bmatrix} \right) = \mathcal{N}\left(0_{2\times 1}, R\right).$$

We assume that labor income risk is uncorrelated with real rate and expected inflation innovations.

This structure delivers a familiar conditionally Gaussian term structure model. The important innovation in this model relative to the literature on mortgage choice is that the market prices of risk λ_t are *time-varying*. The nominal pricing kernel $M^{\$}$ takes the form:

$$\log M_{t+1}^{\$} = -y_t^{\$}(1) - \frac{1}{2}\lambda_t' R \lambda_t - \lambda_t' \varepsilon_{t+1},$$

[11] Brennan and Xia (2002) show that the utility costs induced by incompleteness of the financial market due to unexpected inflation are small. In a previous version of this paper, we have done a numerical, multi-period mortgage choice analysis. We found that unexpected inflation risk did not affect the household's risk-return tradeoff in any meaningful way.

with $\varepsilon_{t+1} = \left[\varepsilon^y_{t+1}, \varepsilon^x_{t+1}\right]'$ and $\lambda_t = [\lambda^y_t, \lambda^x_t]'$. If we were to restrict the prices of risk to be affine, our model would fall in the class of affine term structure models (see Dai and Singleton (2000)), but no such restriction is necessary.

The no-arbitrage price of a two-period zero-coupon bond is:

$$e^{-2y^\$_0(2)} = \mathbb{E}_0\left[M^\$_{t+1} M^\$_{t+2}\right] = e^{-y^\$_0(1) - \mathbb{E}_0\left(y^\$_1(1)\right) + \lambda_0' R\sigma + \frac{1}{2}\sigma' R\sigma},$$

with $\sigma = [\sigma_y, \sigma_x]'$. This equation implies that the long rate equals the average expected future short rate plus a time-varying nominal bond risk premium $\phi^\$$:

$$y^\$_0(2) \;=\; \frac{y^\$_0(1) + \mathbb{E}_0\left(y^\$_1(1)\right)}{2} - \frac{\lambda_0' R\sigma}{2} - \frac{1}{4}\sigma' R\sigma = \frac{y^\$_0(1) + \mathbb{E}_0\left(y^\$_1(1)\right)}{2} + \phi^\$_0(2). \quad (4.6)$$

The long-term nominal bond risk premium $\phi^\$_0(2)$ contains the market price of risk λ_0 and absorbs the Jensen correction term.

4.3.3 Mortgage Pricing

A competitive fringe of mortgage lenders prices ARM and FRM contracts to maximize profit, taking as given the term structure of Treasury interest rates generated by $M^\$$.

Denote the ARM rate at time t by q^{ARM}_t. This is the rate applied to the mortgage payment due in period $t+1$. In each period, the zero-profit condition for the ARM rate satisfies:

$$B = \mathbb{E}_t\left[M^\$_{t+1}\left(q^{ARM}_t + 1\right) B\right] = \left(q^{ARM}_t + 1\right) B P^\$_t(1).$$

This implies that the ARM rate is equal to the one-period nominal short rate, up to an approximation:

$$q^{ARM}_t = P^\$_t(1)^{-1} - 1 \simeq y^\$_t(1).$$

Similarly, the zero-profit condition for the FRM contract stipulates that the present discounted value of the FRM payments must equal the initial loan balance:

$$B = \mathbb{E}_0\left[M^\$_1 q^{FRM}_0 B + M^\$_1 M^\$_2 q^{FRM}_0 B + M^\$_1 M^\$_2 B\right] = q^{FRM}_0 P^\$_0(1) B + \left[q^{FRM}_0 + 1\right] P^\$_0(2) B.$$

Per definition, the nominal interest rate on the FRM is fixed for the duration of the contract. We abstract from the prepayment option for now, but examine its role in Section 4.6.1. The FRM rate, which is a two-period coupon-bearing bond yield, is then equal to:

$$q^{FRM}_0 = \frac{1 - P^\$_0(2)}{P^\$_0(1) + P^\$_0(2)} \simeq \frac{2y^\$_0(2)}{2 - y^\$_0(1) - 2y^\$_0(2)} \simeq y^\$_0(2).$$

The FRM rate is approximately equal to the two-period nominal bond rate.

Our setup embeds two assumptions that merit discussion. The first assumption is that the stochastic discount factor $M^{\$}$ that prices the term structure of interest rates is different from the inter-temporal marginal rate of substitution of the households in section 4.3.1. Without this assumption, mortgage choice would be indeterminate.[12] The second assumption is that we price mortgages as derivatives contracts on the Treasury yield curve. Hence, the same sources that drive time variation in the Treasury yield curve will govern time variation in mortgage rates.

4.3.4 A Household's Mortgage Choice

We now derive the optimal mortgage choice for the household of Section 4.3.1. The crucial difference between an FRM investor and an ARM investor is that the former knows the value of all nominal mortgage payments at time 0, while the latter knows the value of the nominal payments only one period in advance. The risk-averse investor trades off lower expected payments on the ARM against higher variability of the payments. Appendix 4.B computes the life-time utility under the ARM and the FRM contract. It shows that household j prefers the ARM contract over the FRM contract if and only if

$$\begin{aligned} & q_0^{FRM} - q_0^{ARM} + \left(q_0^{FRM} - \mathbb{E}_0\left[q_1^{ARM}\right]\right) e^{-\mathbb{E}_0[x_1]} > \\ & \frac{\gamma_j}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left[\sigma' R \sigma + \left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)^2 \sigma_x^2 - 2\left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)\left(\sigma_x e_2' R \sigma\right)\right] \\ & -\frac{\gamma_j}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left(q_0^{FRM} + 1\right)^2 \sigma_x^2. \end{aligned} \tag{4.7}$$

The left-hand side measures the difference in expected payments on the FRM and the ARM. All else equal, a household prefers an ARM when the expected payments on the FRM are higher than those on the ARM. Appendix 4.B shows that the difference between the expected mortgage payments on the FRM and ARM contracts approximately equals the two-period bond risk premium $\phi_0^{\$}(2)$. This leads to the main empirical prediction of the model: the ARM contract becomes more attractive in periods in which the bond risk premium is high.

The right-hand side of (4.7) measures the risk in the payments, where we recall that γ_j controls risk aversion. The first line arises from the variability of the ARM payments, the

[12] Any equilibrium model of the mortgage market requires a second group of unconstrained investors. Time variation in risk premia could then arise from time-varying risk-sharing opportunities between the constrained and the unconstrained agents, as in Lustig and Van Nieuwerburgh (2006). In their model, the unconstrained agents price the assets at each date and state. Such an environment justifies taking bond prices as given when studying the problem of the constrained investors. Lustig and Van Nieuwerburgh (2006) consider agents with (identical) CRRA preferences. In numerical work, presented in Appendix 4.D, we verify that the same risk-return tradeoff that the constrained households face also hold for CRRA preferences. A full-fledged equilibrium analysis of the mortgage market is beyond the scope of the current paper.

second line represents the variability of the FRM payments. All else equal, a risk-averse household prefers the ARM when the payments on the ARM are less variable than those on the FRM. The risk in the FRM contract is inflation risk (σ_x^2). The balance and the interest payments erode with inflation. The risk in the ARM contract consists of three terms. ARMs are risky because the nominal contract rate adjusts to the nominal short rate each period. The variance of the nominal short rate is $\sigma' R\sigma$. The second term is expected inflation risk, which enters in the same form as in the FRM contract. However, inflation risk is offset by the third term which arises from the positive covariance between expected inflation and the nominal short rate ($\sigma_x e_2' R\sigma$). In low inflation states the mortgage balance erodes only slowly, but the low nominal short rates and ARM payments provide a hedge. The appendix shows that the risk in the ARM is approximately equal to the variability of the real rate (σ_y^2). In sum, the risk-return tradeoff of household j in (4.7), for some generic period t, can be written concisely as:

$$\phi_t^{\$}(2) - \frac{\gamma_j}{2} B\sigma_y^2 + \frac{\gamma_j}{2} B\sigma_x^2 > 0. \tag{4.8}$$

4.3.5 Yield Spread and Long Yield are Poor Proxies

We are the first to suggest the long-term bond risk premium as the determinant of household's mortgage choice. It is the risk premium that is earned on investing in a nominal long-term bond and financing this investment by rolling over a short position in a nominal short-term bond.[13] It is important to emphasize that the long-term bond risk premium is markedly different from both the yield spread and the long-term yield, both of which have been used in the literature to predict mortgage choice.

Using equation (4.6), the difference between the long yield (on the two-period bond) and the short yield (on the one-period bond) can be written as

$$y_0^{\$}(2) - y_0^{\$}(1) \;\; = \;\; \phi_0^{\$}(2) + \frac{\mathbb{E}_0\left(y_1^{\$}(1)\right) - y_0^{\$}(1)}{2}. \tag{4.9}$$

The multi-period equivalent for some generic date t and generic maturity τ is

$$y_t^{\$}(\tau) - y_t^{\$}(1) = \phi_t^{\$}(\tau) + \left(\frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{E}_t\left[y_{t+j-1}^{\$}(1)\right] - y_t^{\$}(1)\right). \tag{4.10}$$

In both expressions, the second term on the right introduces an errors-in-variables problem when the yield spread is used as a proxy for the long-term bond risk premium $\phi_0^{\$}(2)$. This

[13] The strategy holds a τ-period bond until maturity and finances it by rolling over the 1-year bond for τ periods. This definition is different from the one-period bond risk premium in which the long-term bond is held for one period only. Cochrane and Piazzesi (2006) study various definitions of bond risk premia, including ours.

errors-in-variables problem turns out to be so severe that the yield spread has no predictive power for mortgage choice. To understand this further, consider two stark cases. First, in a homoscedastic world with zero risk premia ($\phi_t^{\$}(\tau) = 0$), the yield spread equals the difference between the average expected future short rates and the current short rate. Since long-term bond rates are the average of current and expected future short rates, both the FRM and the ARM investor face the same expected payment stream. The yield spread is completely uninformative about mortgage choice. Second, in a world with constant risk premia, *variations* in the yield spread capture *variations* in deviations between expected future short rates and the current short rate. But again, these variations are priced into both the ARM and the FRM contract. It is only the bond risk premium which affects the mortgage choice for a risk-averse investor. The problem with the yield spread as a measure of the relative desirability of the ARM contract is intuitive: The current short yield is not a good measure for the expected payments on an ARM contract because the short rate exhibits mean reversion which changes expected future payments.

The long yield suffers from a similar errors-in-variables problem:

$$y_0^{\$}(2) \quad = \quad \phi_0^{\$}(2) + \frac{y_0^{\$}(1) + \mathbb{E}_0\left(y_1^{\$}(1)\right)}{2}, . \tag{4.11}$$

where the second term on the right again introduces noise in the predictor of mortgage choice. The problem with the long yield as a measure of the relative desirability of the ARM contract is intuitive: it contains no information on the difference in expected payments between the two contracts. In conclusion, our simple rational mortgage model suggests that both the yield spread and the long-term yield are imperfect predictors of mortgage choice.

4.3.6 Aggregate Mortgage Choice

We aggregate the individual households' mortgage choices to arrive at the ARM share. Define the cutoff risk aversion coefficient that makes a household indifferent between the ARM and FRM contract by:

$$\gamma_t^{\star} \equiv \frac{2\phi_t^{\$}(2)}{B\left(\sigma_y^2 - \sigma_x^2\right)}.$$

Empirically, we find that $\left(\sigma_y^2 - \sigma_x^2\right) > 0$, which guarantees a positive value for the cutoff $\gamma_t^{\star}$. Households that are relatively risk tolerant, with $\gamma_j < \gamma_t^{\star}$, prefer the ARM contract. Because F is the cumulative density function of the risk-aversion distribution, the ARM share is given

by:

$$ARM_t \equiv F(\gamma_t^\star),$$

The complementary fraction of (more risk-averse) households chooses the FRM. The location parameter of the distribution of risk aversion determines the unconditional level of the ARM share. The scale parameter of this distribution drives the sensitivity of aggregate mortgage choice to changes in the bond risk premium. If risk preferences are highly dispersed, the ARM share will be insensitive to changes in the bond risk premium. Conversely, if heterogeneity across households is limited, small changes in the bond risk premium induce large shifts in the ARM share. Hence, the model provides a mapping between the (reduced-form) coefficients of a regression of the ARM share on a constant and the nominal bond risk premium and the two structural parameters that govern the cross-sectional distribution of risk aversion.

4.3.7 Alternative Determinants of Mortgage Choice

Our stylized model of mortgage choice abstracts from several real-life features that are potentially important. Several such features would be straightforward to add to our model, for example stochastic real house prices, a temporary and a permanent component in labor income, and a more general correlation structure between real rate and expected inflation innovations on the one hand and labor income and house prices on the other hand. We could also extend the model to allow for saving in one-period bonds. For realism, we would then impose borrowing constraints along the lines of the life-cycle literature (Cocco, Gomes, and Maenhout (2005)). The models of Campbell and Cocco (2003) and Van Hemert (2006) allow for such features -and more- in the context of a life-cycle model. Campbell and Cocco (2003) show that households with a large mortgage, risky labor income, high risk aversion, a high cost of default, and a low probability of moving are more likely to prefer an FRM contract. In both studies, bond risk premia are assumed to be constant. Our model's sole purpose is understand the link between the term structure of interest rates and both individual and aggregate mortgage choice. We find that the long-term bond risk premium, and not the yield spread or the long yield, is the key determinant of mortgage choice. This is the hypothesis we test empirically in Section 4.4.

4.4 Empirical Results

The main task to render the theory testable is to measure the nominal bond risk premium. The latter is the difference between the current nominal long interest rate and the average

expected future nominal short rate (see (4.6)):

$$\phi_t^{\$}(\tau) = y_t^{\$}(\tau) - \frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{E}_t\left[y_{t+j-1}^{\$}(1)\right]. \tag{4.12}$$

The difficulty resides in measuring the second term on the right, average expected future short rates.

4.4.1 Household Decision Rule

If we assume that households measure expected future short rates by forming simple averages of past short rates, we arrive at the household decision rule $\kappa_t(\rho;\tau)$ of Section 4.2:

$$\begin{aligned} \phi_t^{\$}(\tau) &\simeq y_t^{\$}(\tau) - \frac{1}{12\times\tau}\sum_{s=1}^{\tau\times 12}\left\{\frac{1}{\rho}\sum_{u=0}^{\rho-1} y_{t-u}^{\$}(1)\right\} \\ &= y_t^{\$}(\tau) - \frac{1}{\rho}\sum_{u=0}^{\rho-1} y_{t-u}^{\$}(12) \equiv \kappa_t(\rho;\tau). \end{aligned} \tag{4.13}$$

Equation (4.13) is a model of *adaptive expectations* that only requires knowledge of the current long bond rate, a history of recent short rates (ρ months), and the ability to calculate a simple average. The adaptive expectations scheme delivers a simple proxy $\kappa_t(\rho;\tau)$ for the theoretical bond risk premium $\phi_t^{\$}(\tau)$. Panel A of Figure 4.3 shows the $\tau = 5$- and $\tau = 10$-year time series with a three year look-back, and computed off Treasury interest rates. The two series have a correlation of 92%.[14]

Our main empirical exercise is to regress the ARM share on the nominal bond risk premium. We lag the predictor variable for one month in order to study what changes in this month's risk premium imply for next month's mortgage choice. In addition, the use of lagged regressors mitigates potential endogeneity problems that would arise if mortgage choice affected the term structure of interest rates.[15] The first two rows of Table 4.2 shows the slope coefficient, its Newey-West t-statistic using 12 lags, and the regression R^2 for these

[14]Since we consider look-back periods of up to 5 years, we loose the first 5 years of observations, and the series start in 1989.12. This is the same sample as used in Figures 4.1 and 4.2. We do not extend the sample before 1985.1 for two reasons. First, the interest rates in the early 1980s were dramatically different from those in the period we analyze. As such, we do not consider it to be plausible that households use adaptive expectations and data from the "Volcker regime" to form κ in the first years of our sample. A second and related reason is that Butler, Grullon, and Weston (2006) argue that there is a structural break in bond risk premia in the early 1980s. To avoid any spurious results due to structural breaks, we restrict attention to the period 1985.1-2006.6.

[15]As a robustness check, we have tested for Granger causality. First, we regress the ARM share on its own lag and the lagged bond risk premium; the lagged bond risk premium is statistically significant. Second, we regress the bond risk premium on its lag and the lagged ARM share; the lagged ARM share is statistically insignificant. Therefore, the bond risk premium Granges causes the ARM share, but the reverse is not true.

regressions. Throughout the table, the regressors are normalized by their standard deviation for ease of interpretation. They reinforce the point made in Section 4.2 that the household decision rule is a highly significant predictor of the ARM share. The 5-year (10-year) bond risk premium proxy has a t-statistic of 7.1 (7.5) and explains 71% (68%) of the variation in the ARM share. A one-standard deviation increase in the risk premium increases the ARM share by 7-8 percentage points. This is a large effect since the average ARM share is 28.7%. Intuitively, an FRM holder has to pay the bond risk premium. An increase in the risk premium increases the expected payments on the FRM relative to the ARM, and makes the ARM more attractive.

4.4.2 Forward-Looking Measures

The household decision rule is a proxy for the theoretical bond risk premium when an adaptive expectations scheme is used to form the conditional expectation in equation (4.12). From an academic point of view, there are more conventional ways of measuring average expected future short rates. We study two below: one based on forecasters' expectations and one based on a VAR model.

Forecaster Data

Our forecaster data come from Blue Chip Economic Indicators. Twice per year (March and October), a panel of around 40 forecasters predict the average three-month T-bill rate for the next calendar year, and each of the following four calendar years. They also forecast the average T-bill rate over the ensuing five years. We average the consensus forecast data over the first five, or all ten, years to construct the expected future nominal short rate in (4.12). This delivers a semi-annual time series from 1985 until 2006 for $\tau = 5$ and one for $\tau = 10$. We use linear interpolation of the forecasts to construct monthly series.[16] Combining the 5-year (10-year) T-bond yield with the 5-year (10-year) expected future short rate from Blue Chip delivers the 5-year (10-year) nominal bond risk premium. Panel B of Figure 4.3 shows the 5-year (solid line) and 10-year time series (dashed line); they have a correlation of 94%. We then regress the ARM share on the nominal bond risk premium. The 5-year bond risk premium is a highly significant predictor of the ARM share (Row 3). It has a t-statistic of 3.9, and explains 40% of the variation in the ARM share. A one-standard deviation, or one percentage point, increase in the nominal bond risk premium increases the ARM share by 8.6 percentage points. The results with the 10-year risk premium (Row 4) are comparable. The coefficient has a similar magnitude, a t-statistic of 4.2, and an R^2 of 43%.

[16] The correlations with the ARM share are similar using either semi-annual or monthly data.

VAR Model

A second way to form the forward-looking conditional expectation in equation (4.12) is to use a vector auto-regressive (VAR) term structure model, as in Ang and Piazzesi (2003). The state vector Y contains the 1-year ($y_t^{\$}(1)$), the 5-year ($y_t^{\$}(5)$), and the 10-year nominal yields ($y_t^{\$}(10)$), as well as realized 1-year log inflation ($\pi_t = \log \Pi_t - \log \Pi_{t-1}$). We start the estimation in 1985, near the end of the Volcker period. Our stationary, one-regime model would be unfit to estimate the entire post-war history (see Ang and Bekaert (2005) and Fama (2006)). Estimating the model at monthly frequency gives us a sufficiently many observations (258 months). The VAR(1) structure with the 12-month lag on the right-hand side is parsimonious and delivers plausible long-term expectations.[17] We use the letter u to denote time in months, while t continues to denote time in years. The law of motion for the state is

$$Y_{u+12} = \mu + \Gamma Y_u + \eta_{u+12}, \quad \text{with} \quad \eta_{u+12} \mid \mathcal{I}_u \sim D(0, \Sigma_t), \tag{4.14}$$

with $\mathcal{I}_u$ representing the information at time u. The VAR structure immediately delivers average expected future nominal short rates:

$$\frac{1}{\tau}\mathbb{E}_u\left[\sum_{j=1}^{\tau} y_{u+(12\times(j-1))}(1)\right] = \frac{1}{\tau}e_1'\sum_{j=1}^{\tau}\left\{\left(\sum_{i=1}^{j-1}\Gamma^{i-1}\right)\mu + \Gamma^{j-1}Y_u\right\}. \tag{4.15}$$

Together with the nominal long yield, this delivers our VAR-based measure of the nominal bond risk premium. Panel C of Figure 4.3 shows the 5-year and 10-year time series; they have a correlation of 96%.

Rows 5 and 6 of Table 4.2 show the ARM regression results using the VAR-based 5-year and 10-year bond risk premium. Again, both bond risk premia are highly significant predictors of the ARM share. The t-statistics are 4.2 and 3.9. They explain 32% and 35% of the variation in the ARM share, respectively.[18] The economic magnitude of the slope coefficient is again very close to the one obtained from forecasters and to the one estimated from the household decision rule: In all three cases, a one-standard deviation increase in the risk premium increases the ARM share by about 8 percentage points.

[17] As a robustness check, we considered a VAR(2) model and estimated the model on the basis of quarterly instead of monthly data. The results become even somewhat stronger for a second-order VAR model and we found similar results for quarterly data as for monthly data.

[18] We have also considered and estimated a VAR model with heteroscedastic innovations. In such a model, time variation in the volatility of expected inflation and expected real rates delivers two additional channels for variation in mortgage choice. While both conditional volatilities entered with the predicted sign in the regression, neither was statistically significant. Together, these terms added little explanatory power above the nominal bond risk premium.

The analysis in Section 4.3.6 allows us to interpret the 28% average ARM share and the 8% sensitivity of the ARM share to the bond risk premium in terms of the structural parameters of the model, more precisely the location and scale parameters of the cross-sectional risk aversion distribution. We assume a normal distribution for $\log(\gamma)$ and estimate a mean of 5.0 and a standard deviation of 2.9. The implied median level of risk aversion is 155. Appendix 4.D.4 describes the inference procedure in detail.

In sum, the forward-looking measures and the household rule of thumb deliver quantitatively similar sensitivities of the ARM share to the bond risk premium. This suggests that choosing the "right mortgage at the right time" may require less "financial sophistication" of households than previously thought. As evidenced by the higher R^2 in Rows 1 and 2 compared to Rows 3 to 6, the household decision rule turns out to be the strongest predictor. If the adaptive expectations scheme accurately describes households' behavior, we would expect it to explain more of the variation in households' mortgage choice. We discuss the differential performance of the backward- and forward-looking measures further in Section 4.5.

4.4.3 Alternative Interest Rate Measures

The household decision rule has the appealing feature that it nests two commonly-used predictors of mortgage choice as special cases. First, when $\rho = 1$, we recover the yield spread:

$$\kappa_t(1;\tau) = y_t^{\$}(\tau) - y_t^{\$}(1).$$

The yield spread is the optimal predictor of mortgage choice in our model only if the conditional expectation of future short rates equals the current short rate. This is the case only when short rates follow a random walk. Second, when $\rho \to \infty$, then $\kappa_t(\rho;T)$ converges to the long-term yield in excess of the unconditional expectation of the short rate:

$$\lim_{\rho\to\infty} \kappa_t(\rho;T) = y_t^{\$}(T) - \mathbb{E}\left[y_t^{\$}(12)\right], \tag{4.16}$$

by the law of large numbers.[19] Because the second term is constant, all variation in financial incentives to choose a particular mortgage originates from variation in the long-term yield. This rule is optimal when short rates are constant.

For all cases in between the two extremes, the simple model of adaptive expectations puts some positive and finite weight on average recent short-term yields to form conditional expectations. As Section 4.3.5 argued, this is why both the yield spread and the long yield

[19] This requires a stationarity assumption on the short rates.

suffer from an errors-in-variables problem in the ARM share regressions. To understand this problem, consider the VAR model estimates. They show that the two terms on the right-hand side of (4.10) are *negatively* correlated (-.57 for 5-year and -.54 for 10-year yield). One reason why the correlation between the nominal bond risk premium and the difference between expected future short rates and the current short rate is negative is the following. When expected inflation is high, the inflation risk premium -and hence the nominal bond risk premium- tends to be high. But at the same time, expected future short rates are below the current short rate because inflation is expected to revert back to its long-term mean. This negative correlation makes the yield spread a very noisy proxy for the nominal bond risk premium, and is responsible for the low R^2 in the regression of the ARM share on the yield spread. Indeed, Rows 7 and 8 of Table 4.2 confirm that the lagged yield spread explains less than 1% of the variation in the ARM share in the full sample (1985.1-2006.6). The weak case for the yield spread is also evident in the loan-level data. The second panel of Table 4.1 shows that the yield spread carries a much smaller (normalized) coefficient than the bond risk premium in the top panel, has a much lower t-statistic, and helps classify a lot fewer individual loans correctly.

The long yield suffers from a similar errors-in-variables problem. However, the two terms on the right-hand side of equation (4.11) are positively correlated (.58 for 5-year and .66 for 10-year yield, based on VAR estimation), making the problem less severe. Rows 9 and 10 of Table 4.2 show that the long yield explains 37-39% of the ARM share, with a sensitivity coefficient of around 8.5%. The loan-level analysis in the third panel of Table 4.1 shows that the long yield enters the probit regressions with the wrong sign, substantially reducing the appeal of the long yield as a mortgage choice predictor.

An alternative source of interest rate data comes from the mortgage market. We use the 1-year ARM rate as our measure of the short rate and the 30-year FRM rate as our measure of the long rate (see Appendix 4.A). The household decision rule based on mortgage rate data works well. The regression results in Row 11 are for a two-year look-back period, the horizon that maximizes the correlation with the ARM share in the right panel of Figure 4.2, and deliver an R^2 of 60%. The point estimate of 7.3 is similar to the one from the decision rule based on Treasury rates in Row 1. Row 12 shows similar results for a three-year look-back period. As we did for Treasury yields, we also regress the ARM share on the slope of the yield curve (30-year FRM rate minus 1-year ARM rate) and the long yield (30-year FRM rate). Row 13 shows that the FRM-ARM spread has lower explanatory power than the household decision rule, but much higher explanatory power than the Treasury yield spread. This improvement occurs only because the FRM-ARM spread contains additional

information that is not in the Treasury yield spread.[20] The explanatory power of the FRM rate is similar to that of the long Treasury yield (Row 15 and right panel of Figure 4.2).

The rule-of-thumb that we introduce in Section 4.2 is motivated by the theoretical model in Section 4.3 and provides a way to compute the expectations of future short rates in (4.12). We investigate two additional interest rate-based variables which implement alternative, more ad-hoc, rules-of-thumb. The first rule takes the current FRM rate minus the three-year moving average of FRM rate (row 16 of Table 4.2). The second rule does the same, but for the ARM rate (row 17). The first rule captures the idea behind the popular investment advice of "locking in a low long-term rate while you can". The slope coefficients in the FRM and ARM rule are smaller than what we find for the bond risk premium (6.0 and 3.1) and less precisely measured (t-statistics of 3.7 and 2.4). The R^2 in the two regressions are 22% and 6%, respectively. Both alternative rules perform much worse than the household decision rule of Section 4.4.1, which is guided by theory.

4.5 The Recent Episode and the Inflation Risk Premium

The previous sections showed that all three estimates of the theoretical bond risk premium are positively and significantly related to the FRM-ARM choice. In this section, we investigate the difference between the household decision rule, which shows the strongest relationship and is based on adaptive expectations, and the forecasters- and VAR-based measures, which show a somewhat weaker relationship and are based on forward-looking expectations. Figure 4.4 shows that this difference in performance is especially pronounced after 2004. The figure displays the 10-year rolling-window correlation for each of the three measures with the ARM share. While the rule-of-thumb measure has a stable correlation across sub-samples, the performance of the forecasters-based measure as well as the VAR-based measure drop off steeply around 2004.

The reason for this failure is that the ARM share increased substantially between June 2003 and December 2004 with no commensurate increase in the Blue Chip or VAR risk premia measures. A similarly steep drop-off in correlation occurs for the long yield and for the FRM-ARM rate differential, both of which also performed well in the full sample.

[20] The correlation between the FRM-ARM spread and the 10-1-year government bond yield spread is only 32%. This spread also captures the value of the prepayment option, as well as the lenders' profit margin differential on the FRM and ARM contracts. To get at this additional information, we orthogonalize the FRM-ARM spread to the 10-1 yield spread, and regress the ARM share on the orthogonal component (Row 14). For the full sample, we find a strongly significant effect on the ARM share. Partially this is due to the fact that this orthogonal spread component has a correlation of 60% with the fee differential between an FRM and an ARM contract. It only has a correlation of 16% with the rule-of-thumb risk premium.

We explore two possible explanations for why the ARM share was high in 2004 when the forward-looking bond risk premia were low.

4.5.1 Product Innovation in the ARM Segment

A first potential explanation for the increase in the ARM share between June 2003 and December 2004 is product innovation in the ARM segment of the mortgage markets. An important development was the increased popularity of hybrid mortgages: adjustable-rate mortgages with an initial fixed-rate period.[21] Figure 4.E shows our benchmark measure of the ARM share (solid line) alongside a measure of the ARM share that excludes all hybrid contracts with initial fixed-rate period longer than three years. We label this measure $\widetilde{ARM}$. A substantial fraction of the increase in the ARM share in 2003-05 was due to the rise of hybrids. Under this hypothesis the ARM share went up despite the low bond risk premium because new types of ARM mortgage contracts became available that unlocked the dream of home ownership.[22]

To test this hypothesis, we recompute the rolling correlations for $\widetilde{ARM}$, which excludes the hybrids. The correlation with the forecasters-based measure over the last 10-year window improves from 23% to 48%. The correlation over the longest available sample (since 1992) improves from 44% to 67%. In sum, the recently increased prevalence of the hybrids is part of the explanation. However, it cannot account for the entire story.

4.5.2 Forecast Errors

A second potential explanation is that the forecasters made substantial errors in their predictions of future short rates in recent times. We recall that nominal short rates came down substantially from 6% in 2000 to 1% in June 2003. Our Blue Chip data show that forecasters expected short rates to increase substantially from their 1% level in June 2003. Instead, nominal short rates increased only moderately to 2.2% by December 2004. Forecasters substantially over-estimated future short rates starting in the 2003.6-2004.12 period. As a result, the Blue Chip measure of bond risk premia is too low in that episode, and underestimates the desirability of ARMs.

Forecast errors in nominal rates translate in forecast errors for real rates. This is in particular the case when inflation is relatively stable and therefore easier to forecast. Figure

[21] Starting in 1992, we know the decomposition of the ARM by initial fixed-rate period. We are grateful to James Vickery for making these detailed data available to us.

[22] In addition to the hybrid segment, the sub-prime market segment, which predominantly offers ARM contracts, also grew strongly over that period. However, our ARM sample does not contain this market segment.

4.E shows that the Blue Chip consensus forecast for the average real short rate over the next two years shows large disparities with its realized counterpart. We calculate the average expected future real short rate as the difference between the Blue Chip consensus average expected future nominal short rate and the Blue Chip consensus average expected future inflation rate. We calculate the realized real rate as the difference between the realized nominal rate and expected inflation, which we measure as the one-quarter ahead inflation forecast. The realized average future real short rates are calculated from the realized real rates. Finally, the forecast errors are scaled by the nominal short rate to obtain relative forecasting errors. The figure shows huge forecast errors in the 2000-2003 period, relative to the earlier period. The forecast errors are on the order of 1.25 percentage point per year, about 50-75% of the value of the nominal short rate. These large forecast errors motivate the use of the inflation risk premium, as explained below.

A similar problem arises with the VAR-based bond risk premium. The VAR system also fails to pick up the declining short rates in the 2000-2004 period. It therefore also over-predicts the short rate and underestimates the desirability of ARM contracts.

Filtering Out Forecast Errors Forecast errors in the real rate not only help us identify the problem, they also offer the key to the solution. The nominal bond risk premium in the model of Sections 4.3.1 and 4.3.2 contains compensation for both real rate risk and expected inflation risk:

$$\phi_t^{\$}(\tau) = \phi_t^y(\tau) + \phi_t^x(\tau). \tag{4.17}$$

Similar to the nominal risk premium in (4.12), the real rate risk premium, ϕ_t^y, is the difference between the observed real long rate and the average expected future real short rate:

$$\phi_t^y(\tau) \equiv y_t(\tau) - \frac{1}{\tau}\sum_{j=1}^{\tau} \mathbb{E}_t\left[y_{t+j-1}(1)\right], \tag{4.18}$$

where $y_t(\tau)$ is the real yield of a τ-month real bond at time t. Following Ang and Bekaert (2005), we define the inflation premium at time t, ϕ_t^x, as the difference between long-term nominal yields, long-term real yields, and long-term expected inflation:

$$\phi_t^x(\tau) \equiv y_t^{\$}(\tau) - y_t(\tau) - x_t(\tau). \tag{4.19}$$

where long-term expected inflation is given by:

$$x_t(\tau) \equiv \frac{1}{\tau}\mathbb{E}_t\left[\log \Pi_{t+\tau} - \log \Pi_t\right].$$

A key insight is that both the nominal long yield $y_t^{\$}(\tau)$ and the real long yield $y_t(\tau)$

contain expected future real short rates. Thus, their difference does not. Therefore, their difference zeroes out any forecast errors in expected future real short rates. Equation (4.19) shows that the inflation-risk premium, $\phi_t^x(\tau)$, contains the difference between $y_t^{\$}(\tau) - y_t(\tau)$, and therefore does not suffer from the forecast error problem.[23] In short, one way to correct the nominal bond risk premium for the forecast error is to only use the inflation risk premium component.

Measuring the Inflation Risk Premium To implement equation (4.19), we need a measure of long real yields and a measure of expected future inflation rates. Real yield data are available as of January 1997 when the US Treasury introduced Treasury Inflation-Protected Securities (TIPS). We omit the first six months when liquidity was low, and only a 5-year bond was trading. In what follows, we consider two empirical measures for expected inflation. Our first measure for expected inflation is computed from the same semi-annual Blue Chip long-range consensus forecast data we used for the nominal short rate, using the same method, but using the series for the CPI forecast instead of the nominal short rate.[24] The inflation-risk premium is then obtained by subtracting the real long yield and long-term expected inflation from the nominal long yield, as in (4.19).

Alternatively, we can use the VAR to form expected future inflation rates and thereby the inflation risk premium. We start by constructing the 1-year expected inflation series as a function of the state vector

$$x_t(1) \quad = \quad \mathbb{E}_t\left[\pi_{t+1}\right] = e_4'\mu + e_4'\Gamma Y_t, \tag{4.20}$$

where e_4 denotes the fourth unit vector. Next, we use the VAR structure to determine the τ-year expectations of the average inflation rate in terms of the state variables:

$$\frac{1}{\tau}\mathbb{E}_t\left[\sum_{j=1}^{\tau} e_4' Y_{t+j-1}\right] = \frac{1}{\tau} e_4' \sum_{j=1}^{\tau}\left\{\left(\sum_{i=0}^{j-1}\Gamma^{i-1}\right)\mu + \Gamma^{j-1}Y_t\right\}. \tag{4.21}$$

With the long-term expected inflation from (4.21) in hand, we form the inflation risk premium as the difference between the observed nominal yield, the observed real yield, and expected inflation.

[23]The same Blue Chip forecast data, as well as data from the Survey of Professional Forecasters, indeed show that inflation forecasts do not suffer from the same problem as nominal interest rate forecasts. This is consistent with Ang and Bekaert (2005), who argue that inflation forecasts provide the best predictors of future inflation among a wide set of alternatives.

[24]We have compared the inflation forecasts from Blue Chip with those from the Survey of Professional Forecasters, the Livingston Survey, and the Michigan Survey, and found them to be very close. Ang, Bekaert, and Wei (2006) argue that such survey data provides the best inflation forecasts among a wide array of methods.

Results Figure 4.E shows the inflation risk premium (dashed line) alongside the ARM share (solid line). The inflation risk premium is based on Blue Chip forecast data. Between March 2003 and March 2005 (closest survey dates), the inflation risk premium increased by 1.2 percentage points, or two standard deviations. The nominal bond risk premium, in contrast, only increased only by one standard deviation.

Over the period 1997.7-2006.6, the raw correlation between the ARM share and the 5-year (10-year) inflation risk premium is 84% (82%) for the Blue Chip measure and 80% (78%) for the VAR measure. Finally, we regress the ARM share on the 5-year and 10-year inflation risk premium for the period 1997.7-2006.6. For the Blue Chip measure, we find a point estimate of 6.95 (6.97) for the 5-year (10-year) inflation risk premium. The economic effect is therefore comparable to what we find for the nominal bond risk premium (Section 4.4.2). The coefficient is measured precisely; the t-statistic is 8.0 (7.9). The 5-year (10-year) inflation-risk premium alone explains 66% of the variation (67%) in the ARM share. Likewise, for the VAR-based measure, we find a point estimate of 6.80 (6.40) for the 5-year (10-year) inflation risk premium. The coefficient is measured precisely; the t-statistic equals 8.5 (6.8). The inflation risk premium alone explains 64% of the variation (56%) in the ARM share. We conclude that the inflation risk premium has been a very strong determinant of the ARM share in the last ten years.

In conclusion, at the end of the sample, the forward-looking expectations measures of the bond risk premium suffered from large differences between realized average short rates, and what forecasters or a VAR predicted for these same average short rates. The adaptive expectations scheme of the household decision rule did not suffer from the same problem. This explains why it performed much better in predicting the ARM share in the last part of the sample. The inflation risk premium component of the bond risk premium successfully purges that forecast error from the forward-looking bond risk premium measures. We showed that it is a strong predictor of the ARM share in the 1997.7-2006.6 sample.

4.6 Extensions

In this section, we extend our results along several dimensions. First, we analyze the role of the prepayment option. Second, we revisit the role of financial constraints for mortgage choice. Third, we analyze the robustness of the statistical inference. Finally, we study the role of liquidity in the TIPS market for our results.

4.6.1 Prepayment Option

Sofar the analysis has ignored the prepayment option. In the US, an FRM contract typically has an embedded option which allows the mortgage borrower to pay off the loan at will. We show how the presence of the prepayment option affects mortgage choice within the utility framework of Section 4.3.[25]

FRM Rate With Prepayment A household prefers to prepay at time 1 if the utility derived from the ARM contract exceeds that of the FRM contract. Prepayment entails no costs, but this assumption is easy to relax in our framework. It then immediately follows from comparing the time-1 value function that prepayment is optimal if and only if:

$$q_0^{FRMP} > q_1^{ARM},$$

where the superscript P in q_0^{FRMP} indicates the FRM contract with prepayment. The FRM rate with prepayment satisfies the following zero-profit condition. It stipulates that the present value of mortgage payments the lender receives must equate the initial mortgage balance B:

$$\begin{aligned} B &= \mathbb{E}_0\left[M_1^{\$} q_0^{FRMP} B + I_{\left(q_0^{FRMP} > q_1^{ARM}\right)} M_1^{\$} M_2^{\$} q_1^{ARM} B + I_{\left(q_0^{FRMP} \leq q_1^{ARM}\right)} M_1^{\$} M_2^{\$} q_0^{FRMP} B + M_1^{\$} M_2^{\$} B\right] \\ &= q_0^{FRMP} P_0^{\$}(1) B + \left[q_0^{FRMP} + 1\right] P_0^{\$}(2) B - B\mathbb{E}_0\left[M_1^{\$} M_2^{\$} \max\left\{q_0^{FRMP} - q_1^{ARM}, 0\right\}\right], \end{aligned}$$

where the last term represents the value of the embedded prepayment option held by the household. $I_{(x<y)}$ denotes an indicator function that takes a value of one when $x < y$. This option value satisfies:

$$B\mathbb{E}_0\left[M_1^{\$} M_2^{\$} \max\left\{q_0^{FRMP} - q_1^{ARM}, 0\right\}\right] = B\left(1 + q_0^{FRMP}\right)\left[P_0^{\$}(2)\Phi(d_1) - \frac{1}{1 + q_0^{FRMP}} P_0^{\$}(1)\Phi(d_2)\right],$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution, and the expressions for d_1 and d_2 are provided in Appendix 4.C. The second step is an application of the Black and Scholes (1973) formula and is spelled out in Appendix 4.C as well (See also Merton (1973) and Jamshidian (1989)). The household has $B\left(1 + q_0^{FRMP}\right)$ European call options on a two-period bond with expiration date $t = 1$ (when it becomes a one-year bond with price $P_1^{\$}(1) = 1/(1+q_1^{ARM})$), and with an exercise price of $1/(1+q_0^{FRMP})$. Substituting the option

[25] We contribute to the large literature on rational prepayment models, e.g., Dunn and McConnell (1981), Stanton and Wallace (1998), Longstaff (2005), and Pliska (2006), by adding time variation in risk premia. Other studies consider reduced-form models that can accommodate slow prepayment (e.g., Schwartz and Torous (1989), Stanton (1995), Boudoukh, Whitelaw, Richardson, and Stanton (1997), and Schwartz (2007)).

value into the zero-profit condition we get:

$$B = \left(q_0^{FRMP} + \Phi\left(d_2\right)\right) P_0^{\$}\left(1\right) B + \left[q_0^{FRMP} + 1\right] P_0^{\$}\left(2\right) B \left(1 - \Phi\left(d_1\right)\right).$$

The mortgage balance equals the sum of (i) the (discounted) payments at time $t = 1$, a certain interest payment and a principal payment with risk-adjusted probability $\Phi\left(d_2\right)$, and (ii) the (discounted) payments at time $t = 2$, when both interest and principal payments are received with risk-adjusted probability $1 - \Phi\left(d_1\right)$. The no-arbitrage rate q_0^{FRMP} on an FRM with prepayment solves the fixed-point problem:

$$q_0^{FRMP} = \frac{1 - \left(1 - \Phi\left(d_1\right)\right) P_0^{\$}\left(2\right) - \Phi\left(d_2\right) P_0^{\$}\left(1\right)}{P_0^{\$}\left(1\right) + \left(1 - \Phi\left(d_1\right)\right) P_0^{\$}\left(2\right)},$$

which cannot be solved for analytically as q_0^{FRMP} appears in d_1 and d_2 on the right-hand side. For $\Phi\left(d_1\right) = \Phi\left(d_2\right) = 1$, prepayment is certain, and we retrieve the expression for the year-one ARM rate, q_0^{ARM}. For $\Phi\left(d_1\right) = \Phi\left(d_2\right) = 0$, prepayment occurs with zero probability, and we obtain the expression for the FRM without prepayment, q_0^{FRM}.

This framework clarifies the relationship between time-varying bond risk premia and the price of the prepayment option. The bond risk premium goes up when the price of interest rate risk goes down. But a decrease in the price of interest rate risk makes prepayment less likely under the risk-neutral distribution. This is because the risk-neutral distribution shifts to the right and makes low interest rate states, where prepayment occurs, less likely. Therefore, the price of the prepayment option is decreasing in the bond risk premium.

Reduced Sensitivity A fixed-rate mortgage *without* prepayment option is a coupon-bearing nominal bond, issued by the borrower and held by the lender.[26] An FRM *with* prepayment option resembles a callable bond: the borrower has the right to prepay the outstanding mortgage debt at any point in time. The price sensitivity of a callable bond to interest rate shocks differs from that of a regular bond. This is illustrated in Figure 4.E. We use the bond pricing setup of Section 4.3.2 and set $\mu_y = \mu_x = 2\%$, $\rho_y = \rho_x = 0.5$, $\rho_{xy} = 0$, $\sigma_y = \sigma_x = 2\%$, and $\lambda_0 = [-0.4, -0.4]'$. These values imply a two-period nominal bond risk premium of $\phi_0^{\$}\left(2\right) = 0.78\%$. We vary the short rate at time zero, $y_0^{\$}\left(1\right) = y_0\left(1\right) + x_0$, assuming $y_0(1) = x_0$. The callable bond can be called at time one with exercise price of 0.96 (per dollar face value). The non-callable bond price is decreasing and convex in the nominal interest rate. The callable bond price is also decreasing in the nominal interest

[26] This analogy is exact for an interest-only mortgage. When the mortgage balance is paid off during the contractual period (amortizing), the loan can be thought of as a portfolio of bonds with maturities equal to the dates on which the down-payments occur. Acharya and Carpenter (2002) discuss the valuation of callable, defaultable bonds.

rate, but, the relationship becomes concave when the call option is in the money ("negative convexity"). This means that the callable bond has positive, but diminished exposure to nominal interest-rate risk.

Utility Implications of the Prepayment Option Next, we study how the prepayment option affects the relationship between the bond risk premium and the ARM-FRM utility differential. We use the same term-structure variables as in Figure 4.E, but vary the market prices of risk λ_0. We maintain the assumption of equal prices of inflation risk and real interest rate risk, and fix the initial real interest and inflation rate at their unconditional means, i.e. $y_0(1) = \mu_y$ and $x_0 = \mu_x$. We assume the investor has a mortgage balance and house size normalized to 1, constant real labor income of 0.41, and a risk aversion coefficient $\gamma = 10$. Figure 4.E plots the difference between the lifetime utility from the ARM contract and the lifetime utility from the FRM contract. The solid line depicts the case without prepayment option; the dashed line plots the utility difference when the FRM has the prepayment option. No approximations are used for this exercise. The utility difference is increasing in the bond risk premium, both with and without prepayment option. However, the sensitivity of the utility difference to changes in the bond risk premium is somewhat reduced in presence of a prepayment option. This is consistent with the fact that a callable bond has diminished interest rate exposure and therefore contains a lower bond risk premium than a non-callable bond. This shows that our main result, a positive relationship between the utility difference of an ARM and an FRM contract and the nominal bond risk premium, goes through.

4.6.2 Financial Constraints

One alternative hypothesis is that there is a group of financially-constrained households which postpones the purchase of a house until the ARM rate is sufficiently low to qualify for a mortgage loan. Under this alternative hypothesis, the time series variation in the dollar volume of ARMs would drive the variation in the ARM share, and the dollar volume in FRMs would be relatively constant. Figure 4.10 plots the dollar volume of ARM and FRM mortgage originations for the entire U.S. market, scaled by the overall size of the mortgage market. The data are compiled by OFHEO. It shows that there are large year-on-year fluctuations in both the ARM and the FRM market segment. This dispels the hypothesis that the variation in the ARM share over the last 20 years is driven by fluctuations in participation in the ARM segment.

Loan-level data provide arguably the best laboratory to test the importance of financial constraints. As we showed in Section 4.2 and Table 4.1, loan balance, FICO score, and LTV ratio were all significant predictors of the probability of choosing an ARM. However, they

did not drive out the bond risk premium. Rather, the bond risk premium is economically the stronger determinant of mortgage choice based on the size of its coefficient, its t-statistic, and the number of correctly classified loans. The financial constraint variables did not add any explanatory power. This is a powerful result given that the balance, FICO score, and LTV ratio are cross-sectional variables, while the bond risk premium is a time series variable.

There is substantial cross-state variation in mortgage choice in the US. In 2006, the ARM share was above 40% in California, but less than 10% in Connecticut. The loan-level data set is large enough to investigate the relationship between the ARM share on the bond risk premium state-by-state. Interestingly, the size of the probit coefficient on the bond risk premium and its t-statistic are rather similar across states. So while the level of the ARM share may be a function of financial constraint-type variables such as the median house price, we find a strong positive covariation between the bond risk premium and the ARM share for all states.[27]

The importance of the yield spread as a predictor of mortgage choice also relates to the role of financial constraints. Section 4.4.3 showed that the yield spread did not display a strong co-movement with the ARM share. We argued that the yield spread not only captures the bond risk premium, but also deviations of expected future short rates from current short rates, causing the problem. However, when a household is perfectly impatient and only cares about consumption in the current period ($\beta = 0$), only the current period's differential between the long-term and the short-term interest rate matters. The same is true if a household plans to move in the current year.[28] The multi-period model of Appendix 4.D allows us to investigate the quantitative role of the time discount factor and the moving rate for mortgage choice. Our conclusions are that for conventional values of the time discount factor or the moving rate, it is the bond risk premium which matters. Finally, we have investigated the extent to which the yield spread affects mortgage choice in the data, over and above the risk premium. In a multiple time series regression of the ARM share on the risk premium and the yield spread, the latter was typically not significant. Its sign flips across specifications, its t-statistic is low, and it does not contribute to the R^2 of the regression, beyond the effect of the risk premium. In the loan-level data, adding the yield spread to the probit regression with the bond risk premium, the loan balance, FICO score, LTV ratio, and the regional dummies only strengthens the effect of the bond risk premium (Row 6 of Table 4.1). While the yield spread is highly significant in this regression, it does not seem

[27] We also investigated the effect of the aggregate loan-to-value ratio, aggregate house price-income, and house price-rent ratios on the ARM share, but found no relationship.

[28] Mobility in and of itself is an unlikely candidate to explain variation in the ARM share. Current Population Survey data for 1948-2004 from the US Census show that the average annual (monthly) moving rate is 18.1% (1.27%), and the out-of-county moving rate is 6.2% (1.16%). Moreover, these moving rates show no systematic variation over time.

to be the case that its explanatory power captures the effect of financial constraints. The coefficients and significance level of the loan balance, FICO score, and LTV ratio are not diminished. Put differently, adding the yield spread to the bond risk premium has stronger effects than adding these three loan characteristics.

We conclude that, first, the bond risk premium is a powerful predictor of mortgage choice in these loan-level data. Second, while measures of financial constraints certainly enter significantly in these regressions, both their economic and statistical effect on mortgage choice is smaller.

4.6.3 Persistence of Regressor

In contrast to the bond risk premium, most term structure variables do not explain much of the variation in the ARM share (Table 4.2). This is especially true in the last ten years of our sample, when the inflation risk premium has strong explanatory power (Section 4.5), but the real yield or the FRM-ARM rate differential do not. This suggests that our results for the risk premium are not simply an artifact of regressing a persistent regressand on a persistent regressor, because many of the other term structure variables are at least as persistent.[29] To further investigate this issue, we conduct a block-bootstrap exercise, drawing 10,000 times with replacement 12-month blocks of innovations from an augmented VAR. The latter consists of the four equations of the VAR of Section 4.4.2, and is augmented with an equation for the ARM share. The ARM share equation is allowed to depend on the four lagged VAR elements, as well as on its own lag. The lagged ARM share itself does not affect the VAR elements. The bootstrap estimate recovers the point estimate (no bias), and it leads to a confidence interval that is narrower (6.40) than the Newey-West confidence interval we use in the main text (8.24), but wider than an OLS confidence interval (3.73). We conclude that the Newey-West standard errors we report are conservative.

One further robustness check we performed is to regress quarterly changes in the ARM share (between periods t and $t+3$) on changes in the term structure variables of the benchmark regression specification (between periods $t-1$ and t). We continue to find a positive and strongly significant effect of the risk premium on the ARM share (t-statistic around 5). The effect of a change in the bond risk premium is similar to the one estimated from the level regressions: a one percentage point increase in the bond risk premium leads to a 10 percentage point increase in the ARM share over the next quarter. The R^2 of the regression in changes is obviously lower, but still substantial. For the 5-year (10-year) risk premium

[29] The ARM share itself is not that persistent. Its annual autocorrelation is 30%, compared to 76% for the one-year nominal interest rate. An AR(1) at an annual frequency only explains 8.8% of the variation in the ARM share.

based on the VAR, it is 12% (18%), for the forecaster measure it is 25% (30%), and for the rule-of-thumb it is 26% (27%).

4.6.4 Liquidity and the TIPS Market

The results in Section 4.5, which use the inflation risk premium, are based on TIPS data. The TIPS markets suffered from liquidity problems during the first years of operation, which may have introduced a liquidity premium in TIPS yields (see Shen and Corning (2001) and Jarrow and Yildirim (2003)). A liquidity premium is likely to induce a downward bias in the inflation risk premium. As long as this bias does not systematically covary with the ARM share, it operates as an innocuous level effect and adds measurement error.

To rule out the possibility that our inflation risk premium results are driven by liquidity premia, we use real yield data backed out from the term structure model of Ang and Bekaert (2005) instead of the TIPS yields. We treat the real yields as observed, and use them to construct the inflation risk premium.[30] Since the Ang-Bekaert-Wei data are quarterly (1985.IV-2004.IV), we construct the quarterly ARM share as the simple average of the three monthly ARM share observations in that quarter. We then regress the quarterly ARM share on the one-quarter lagged inflation and real rate risk premium. We find that both components of the nominal bond risk premium, the inflation-risk premium, and the real rate risk premium, enter with a positive sign. This is consistent with the theoretical model developed in Section 4.3. Both coefficients are statistically significant: The Newey-West t-statistic on the inflation risk premium is 3.90 and the t-statistic on the real rate risk premium is 2.12. The regression R-squared is 53%.

As a final robustness check, we repeated our regressions using only TIPS data after 1999.1, after the initial period of illiquidity. We found very similar results to those based on data starting in 1997.7. This suggests that liquidity problems in TIPS markets may have affected the inflation-risk premium, but this does not significantly affect our results. We conclude that our results are robust to using alternative real yield data.

4.7 Conclusion

We have shown that the time variation in the nominal risk premium on a long-term nominal bond can explain a large fraction of the variation in the share of newly-originated mortgages that are of the adjustable-rate type. Thinking of fixed-rate mortgages as a short position in long-term bonds and adjustable-rate mortgages as rolling over a short position in short-term bonds implies that fixed-rate mortgage holders are paying a nominal bond risk premium. The

[30] We thank Andrew Ang for making these data available to us.

higher the bond risk premium, the more expensive the FRM, and the higher the ARM share. Our results are consistent across three different methods of computing bond risk premia. We used forecasters' expectations, a VAR-model, and a simple adaptive expectation scheme, or "household decision rule". This last measure explains 70% of the variation in the ARM share. Other term structure variables, such as the slope of the yield curve, have much lower explanatory power for the ARM share.

For all three measures of the bond risk premium, a one standard deviation increase leads to an eight percentage point increase in the ARM share. Studying these different risk premium measures also reveals interesting differences. In the last ten years of our sample, only the household decision rule continues to predict the ARM share. We track the poorer performance of the forecasters-based measure down to large forecast errors in future short rates. We show that these forecast errors are not present in the inflation risk premium component of the bond risk premium. We use real yield data and inflation forecasts to construct the inflation risk premium and show that it has strong predictive power for the ARM share. This exercise lends further credibility to the bond risk premium as the relevant term structure variable for mortgage choice.

In a previous version of the paper, we have also studied mortgage choice in the UK. Fixed rate mortgages are a lot less prevalent in the UK than in the US, and only a recent addition to the market. While the maturity choice may be somewhat less relevant, we still found a similar positive covariation between the ARM share and the bond risk premium.

Taken together, our findings suggest that households may be making close-to-optimal mortgage choice decisions. Capturing the relevant time variation in bond risk premia is feasible by using a simple rule. This paper contributes to the growing household finance literature (Campbell (2006)), which debates the extent to which households make rational investment decisions. Given the importance of the house in the median household's portfolio and the prevalence of mortgages to finance the house, the problem of mortgage origination deserves a prominent place in this debate.

4.A Data

Aggregate Time Series Data for ARM Share Our baseline data series is from the Federal Housing Financing Board. It is based on the Monthly Interest Rate Survey (MIRS), a survey sent out to mortgage lenders. The monthly data start in 1985.1 and run until 2006.6, and we label this series $\{ARM_t\}$. Major lenders are asked to report the terms and conditions on all conventional, single-family, fully-amortizing, purchase-money loans closed the last five working days of the month. The data thus excludes FHA-insured and VA-guaranteed mortgages, refinancing loans, and balloon loans. The data for our last sample month, June 2006, are based on 21,801 reported loans from 74 lenders, representing savings associations, mortgage companies, commercial banks, and mutual savings banks. The data are weighted to reflect the shares of mortgage lending by lender size and lender type as reported in the latest release of the Federal Reserve Board's Home Mortgage Disclosure Act data. They are available at `http://www.fhfb.gov/Default.aspx`.

These MIRS data include only new house purchases (for both newly-constructed homes and existing homes), not refinancings. Freddie Mac publishes a monthly index of the share of refinancings in mortgage originations. The average refi share over the 1987.1-2007.1 period is 39.3%. So, purchase-money loans accounts for approximately 60% of the mortgage flow. The sample consists predominantly of conforming loans, only a very small fraction is jumbo mortgages. The ARM share for jumbos in the MIRS sample is much higher on average, but has a 70% correlation with the conforming loans in the sample. While the data do not permit precise statements about the representativeness of the MIRS sample, its ARM share has a correlation of 94% with the ARM share in the Inside Mortgage Finance data. The comparison is for annual data between 1990 and 2006, the longest available sample. We thank Nancy Wallace for making the IMF data available to us.

There is an alternative source of monthly ARM share data available from Freddie-Mac, based on the Primary Mortgage Market Survey. This survey goes out to 125 lenders. The share is constructed based on the dollar volume of conventional mortgage originations within the 1-unit Freddie Mac loan limit as reported under the Home Mortgage Disclosure Act (HMDA) for 2004. Given that Freddie Mac also publishes the aforementioned refinancing share of originations based on the same Primary Mortgage Market Survey, it appears that this series includes not only purchase mortgages but also refinancings. This series is available from 1995.1 and has a correlation with our benchmark measure of 90%.

Loan-level Mortgage Data We explore a new data set which contains information on 911,000 loans from a large mortgage trustee for mortgage-backed security special purpose vehicles. It contains data from many of the largest mortgage lenders such as Aames Capital, Bank of America, Citi Mortgage, Countrywide, Indymac, Option One, Ownit, Wells Fargo, Washington Mutual. We use information on the loan type, the loan origination year and month, the balance, the loan-to-value ratio, the FICO score, and the contract rate at origination. We also have geographic information on the region of origination. We merge these data with our bond risk premium and interest rate variables, with matching based on month of origination. While the sample spans 1994-2007, 95% of mortgage contracts are originated between 2000 and 2005.

Treasury and Mortgage Yields and Inflation Monthly nominal yield data are obtained from the Federal Reserve Bank of New York. They are available at `http://www.federalreserve.gov/pubs/feds/2006`. We use the 1-year ARM rate as our measure of the short mortgage rate and the 30-year FRM rate as our measure of the long-term mortgage rate. We use the effective rate data from the Federal Housing Financing Board, Table 23. The effective rate adjusts the contractual rate for the discounted value of initial fees and

charges. The FRM-ARM spreads with and without fees have a correlation of .998. The inflation rate is based on the monthly Consumer Price Index for for all urban consumers from the Bureau of Labor Statistics. The inflation data are available at `http://www.bls.gov`. Real yield data are available as of January 1997 when the US Treasury introduced Treasury Inflation-Protected Securities (TIPS). The real yield data are available from McCulloch at `http://www.econ.ohio-state.edu/jhm/ts/ts.html`.

4.B Risk-Return Tradeoff

This appendix computes the expected utility from time-1 and time-2 consumption for each of the contracts. We first compute the utility without log transformation, and only at the end, when comparing the two mortgage contracts, reintroduce this log transformation.

Utility from time-1 consumption The (exponent of) utility from time-1 consumption on the FRM contract is:

$$\begin{aligned}\mathbb{E}_0\left(e^{-\beta-\gamma\frac{C_1}{\Pi_1}}\right)=\mathbb{E}_0\left(e^{-\beta-\gamma\left[\frac{L_1^{\$}-q_0^{FRM\,B}}{\Pi_1}\right]}\right) &= \mathbb{E}_0\left(e^{-\beta-\gamma\left[L_1-\frac{q_0^{FRM\,B}}{\Pi_1}\right]}\right)\\ &= e^{-\beta-\gamma\left(\mathbb{E}_0(L_1)-\frac{\gamma\sigma_L^2}{2}-\frac{q_0^{FRM\,B}}{\Pi_1}\right)}.\end{aligned}$$

For the ARM contract it is:

$$\mathbb{E}_0\left(e^{-\beta-\gamma\frac{C_1}{\Pi_1}}\right)=\mathbb{E}_0\left(e^{-\beta-\gamma\left[\frac{L_1^{\$}-q_0^{ARM\,B}}{\Pi_1}\right]}\right)=e^{-\beta-\gamma\left(\mathbb{E}_0(L_1)-\frac{\gamma\sigma_L^2}{2}-\frac{q_0^{ARM\,B}}{\Pi_1}\right)}.$$

Utility from time-2 consumption Under the FRM, the time-1 value of the time-2 utility equals:

$$\mathbb{E}_1\left[e^{-2\beta-\gamma\frac{C_2}{\Pi_2}}\right]=e^{-2\beta-\gamma\left(H_2+\mathbb{E}_1[L_2]-\frac{\gamma\sigma_L^2}{2}-\frac{\left(q_0^{FRM}+1\right)B}{\Pi_2}\right)},$$

using the same argument as in the period-1 utility calculations.

Next, we calculate the time-0 utility of this time-2 utility:

$$\begin{aligned}\mathbb{E}_0\left[e^{-2\beta-\gamma\frac{C_2}{\Pi_2}}\right] &= \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-\left(q_0^{FRM}+1\right)Be^{-x_0-x_1}\right)}\right]\\ &\simeq \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-\left(q_0^{FRM}+1\right)Be^{-x_0-\mathbb{E}_0[x_1]}(1-(x_1-\mathbb{E}_0[x_1]))\right)}\right]\\ &= \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-\left(q_0^{FRM}+1\right)Be^{-x_0-\mathbb{E}_0[x_1]}(1-\sigma_x\varepsilon_1^x)\right)}\right]\\ &= e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]-\left(q_0^{FRM}+1\right)Be^{-x_0-\mathbb{E}_0[x_1]}\right)+\frac{\gamma^2}{2}\left(\left(1+\rho_L^2\right)\sigma_L^2+\left(q_0^{FRM}+1\right)^2B^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma_x^2\right)}.\end{aligned}$$

In these steps, we used:

$$\begin{aligned}
\Pi_2 &= \Pi_1 e^{x_1}, \quad \Pi_1 = e^{x_0}, \\
\mathbb{E}_1(L_2) &= \mu_L + \rho_L(L_1 - \mu_L) = \mu_L + \rho_L^2(L_0 - \mu_L) + \rho_L \sigma_L \varepsilon_1^L = \mathbb{E}_0(L_2) + \rho_L \sigma_L \varepsilon_1^L, \\
e^{-x_1} &\simeq e^{-\mathbb{E}_0(x_1)} - e^{-\mathbb{E}_0(x_1)}[x_1 - \mathbb{E}_0(x_1)].
\end{aligned}$$

For the ARM contract, the time-1 value of the time-2 utility equals:

$$\mathbb{E}_1\left[e^{-2\beta-\gamma\frac{C_2}{\Pi_2}}\right] = e^{-2\beta-\gamma\left(H_2+\mathbb{E}_1(L_2)-\frac{\gamma\sigma_L^2}{2}-\frac{\left(1+q_1^{ARM}\right)B}{\Pi_2}\right)}.$$

Then for the time-0 value function, it holds:

$$\begin{aligned}
\mathbb{E}_0\left[e^{-2\beta-\gamma\frac{C_2}{\Pi_2}}\right] &\simeq \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-\left(1+q_1^{ARM}\right)Be^{-x_0-\mathbb{E}_0[x_1]}(1-\sigma_x\varepsilon_1^x)\right)}\right] \\
&= \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-B\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1+q_1^{ARM}-\mathbb{E}_0\left[q_1^{ARM}\right]\right)e^{-x_0-\mathbb{E}_0[x_1]}(1-\sigma_x\varepsilon_1^x)\right)}\right] \\
&\simeq \mathbb{E}_0\left[e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]+\rho_L\sigma_L\varepsilon_1^L-\frac{\gamma\sigma_L^2}{2}-B\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1+\sigma'\varepsilon_1\right)e^{-x_0-\mathbb{E}_0[x_1]}(1-\sigma_x\varepsilon_1^x)\right)}\right] \\
&= e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]-B\left(\mathbb{E}_0\left(q_1^{ARM}\right)+1\right)e^{-x_0-\mathbb{E}_0(x_1)}\right)+\frac{\gamma^2}{2}\left[\left(1+\rho_L^2\right)\sigma_L^2\right]} \times \\
&\quad \mathbb{E}_0\left[e^{-\gamma B\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)e^{-x_0-\mathbb{E}_0[x_1]}\sigma_x\varepsilon_1^x+\gamma Be^{-x_0-\mathbb{E}_0[x_1]}\sigma'\varepsilon_1-\gamma Be^{-x_0-\mathbb{E}_0[x_1]}\left(\sigma'\varepsilon_1\right)\sigma_x\varepsilon_1^x}\right] \\
&\simeq e^{-2\beta-\gamma\left(H_2+\mathbb{E}_0[L_2]-B\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)e^{-x_0-\mathbb{E}_0[x_1]}\right)} \times \\
&\quad e^{\frac{\gamma^2}{2}\left[\left(1+\rho_L^2\right)\sigma_L^2+B^2\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma_x^2+B^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma'R\sigma-2B^2\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)e^{-2x_0-2\mathbb{E}_0[x_1]}\left(\sigma_xe_2'R\sigma\right)\right.}
\end{aligned}$$

The last approximation assumes that $\gamma e^{-x_0-\mathbb{E}_0(x_1)}(\sigma'\varepsilon_1)\sigma_x\varepsilon_1^x$ is zero (a shock times a shock). $(\sigma_x e_2' R\sigma)$ is the covariance of x and $y^{\$}$, where we defined $e_2 = [0,1]'$. In the third line of the approximation, we use $q_1^{ARM} \simeq y_1^{\$}(1)$.

Now we reintroduce the log transformation to the exponential preferences. Households prefer the ARM if and only if the life-time utility of the ARM contract exceeds that of the FRM contract:

$$\begin{aligned}
&\beta + \gamma\left(\mathbb{E}_0(L_1) - \frac{\gamma\sigma_L^2}{2} - \frac{q_0^{ARM}B}{\Pi_1}\right) \\
&+2\beta + \gamma\left(H_2 + \mathbb{E}_0[L_2] - B\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)e^{-x_0-\mathbb{E}_0[x_1]}\right) \\
&-\frac{\gamma^2}{2}\left[\begin{array}{l}\left(1+\rho_L^2\right)\sigma_L^2 + B^2\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma_x^2 \\ +B^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma'R\sigma - 2B^2\left(\mathbb{E}_0\left[q_1^{ARM}\right]+1\right)e^{-2x_0-2\mathbb{E}_0[x_1]}\left(\sigma_xe_2'R\sigma\right)\end{array}\right) \\
>& \\
&\beta + \gamma\left(\mathbb{E}_0(L_1) - \frac{\gamma\sigma_L^2}{2} - \frac{q_0^{FRM}B}{\Pi_1}\right) \\
&+2\beta + \gamma\left(H_2 + \mathbb{E}_0[L_2] - \left(q_0^{FRM}+1\right)Be^{-x_0-\mathbb{E}_0[x_1]}\right) - \frac{\gamma^2}{2}\left(\left(1+\rho_L^2\right)\sigma_L^2 + \left(q_0^{FRM}+1\right)^2B^2e^{-2x_0-2\mathbb{E}_0[x_1]}\sigma_x^2\right).
\end{aligned}$$

This simplifies to:

$$\begin{aligned} & q_0^{FRM} - q_0^{ARM} + \left(q_0^{FRM} - \mathbb{E}_0\left[q_1^{ARM}\right]\right) e^{-\mathbb{E}_0[x_1]} \\ > \\ & \frac{\gamma}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left[\sigma' R \sigma + \left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)^2 \sigma_x^2 - 2\left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)\left(\sigma_x e_2' R \sigma\right)\right] \\ & -\frac{\gamma}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left(q_0^{FRM} + 1\right)^2 \sigma_x^2. \end{aligned}$$

Simplifying Expressions The first term on the right-hand side of the inequality, i.e., the risk induced by the ARM contract, can be rewritten as:

$$\begin{aligned} & \frac{\gamma}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left[\sigma' R \sigma + \left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)^2 \sigma_x^2 - 2\left(\mathbb{E}_0\left[q_1^{ARM}\right] + 1\right)\left(\sigma_x e_2' R \sigma\right)\right] \\ = \; & \frac{\gamma}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \left[\sigma_y^2 - 2\sigma_x \sigma_y \rho_{xy} \mathbb{E}_0\left[q_1^{ARM}\right] + \mathbb{E}_0\left[q_1^{ARM}\right]^2 \sigma_x^2\right] \\ \simeq \; & \frac{\gamma}{2} B e^{-x_0 - 2\mathbb{E}_0[x_1]} \sigma_y^2, \end{aligned}$$

in which we use that $2\sigma_x \sigma_y \rho_{xy} \mathbb{E}_0\left[q_1^{ARM}\right]$ and $\mathbb{E}_0\left[q_1^{ARM}\right]^2 \sigma_x^2$ are an order of magnitude smaller than σ_y^2, which motivates the approximation in the third line. This in turn implies that the ARM contract primarily carries real rate risk, while, in contrast, the FRM contract carries only inflation risk. This is the risk-return trade-off discussed in the main text.

Ignoring the $e^{-\mathbb{E}_0[x_1]}$ inflation term, the left-hand side of above inequality is the difference in expected nominal payments per dollar mortgage balance. We have:

$$2q_0^{FRM} - q_0^{ARM} - \mathbb{E}_0\left(q_1^{ARM}\right) \simeq 2y_0^{\$}(2) - y_0^{\$}(1) - \mathbb{E}_0\left[y_1^{\$}(1)\right] = 2\phi_0^{\$}(2)$$

where we use the approximations of Section 4.3.3.

4.C Derivation of the Prepayment Option Formula

The value of the prepayment option is given by:

$$\begin{aligned} B\mathbb{E}_0\left[M_1^{\$} M_2^{\$} \max\left\{\left(q_0^{FRMP} - q_1^{ARM}\right), 0\right\}\right] &= B\mathbb{E}_0\left[\mathbb{E}_1\left[M_1^{\$} M_2^{\$} \max\left\{\left(q_0^{FRMP} - q_1^{ARM}\right), 0\right\}\right]\right] \\ &= B\mathbb{E}_0\left[M_1^{\$} \max\left\{\left(q_0^{FRMP} - q_1^{ARM}\right) P_1^{\$}(1), 0\right\}\right] \\ &= B\mathbb{E}_0\left[M_1^{\$} \max\left\{\left(1 + q_0^{FRMP} - P_1^{\$}(1)^{-1}\right) P_1^{\$}(1), 0\right\}\right] \\ &= B\mathbb{E}_0\left[M_1^{\$} \max\left\{\left(\left(1 + q_0^{FRMP}\right) P_1^{\$}(1) - 1\right), 0\right\}\right] \\ &= B\left(1 + q_0^{FRMP}\right)\mathbb{E}_0\left[M_1^{\$} \max\left\{\left(P_1^{\$}(1) - \frac{1}{1 + q_0^{FRMP}}\right), 0\right\}\right] \end{aligned}$$

where we use that $q_1^{ARM} = P_1^{\$}(1)^{-1} - 1$. The pricing kernel and the one-year bond price at time $t = 1$ are given by:

$$\begin{aligned} M_1^{\$} &= e^{-y_0^{\$}(1) - \frac{1}{2}\lambda_0' R \lambda_0 - \lambda_0' \varepsilon_1} \\ P_1^{\$}(1) &= e^{-y_1^{\$}(1)} = e^{-\mathbb{E}_0\left[y_1^{\$}(1)\right] - \sigma' \varepsilon_1} \end{aligned}$$

We project the innovation to the pricing kernel on the innovation to the nominal short rate:

$$\begin{aligned}\eta_1 &\equiv \sigma'\varepsilon_1 \\ \eta_2 &\equiv \lambda_0'\varepsilon_1 - \frac{\mathrm{Cov}\left(\eta_1, \lambda_0'\varepsilon_1\right)}{\mathrm{Var}\left(\eta_1\right)}\eta_1 = \lambda_0'\varepsilon_1 - \frac{\sigma' R\lambda_0}{\sigma' R\sigma}\eta_1\end{aligned}$$

with η_1 and η_2 orthogonal and variances given by:

$$\mathrm{Var}\left[\eta_1\right] = \sigma' R\sigma, \quad \mathrm{Var}\left[\eta_2\right] = \lambda_0' R\lambda_0 - \frac{\left(\sigma' R\lambda_0\right)^2}{\sigma' R\sigma}$$

We first solve for the value of one call option for a general exercise price K, denoted by $C_0\left(K\right)$:

$$\begin{aligned}C_0\left(K\right) &= \mathbb{E}_0\left[M_1^{\$}\max\left\{\left(P_1^{\$}\left(1\right)-K\right),0\right\}\right] \\ &= \mathbb{E}_0\left[e^{-y_0^{\$}(1)-\frac{1}{2}\lambda_0' R\lambda_0-\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\eta_1-\eta_2}\max\left\{\left(e^{-\mathbb{E}_0\left[y_1^{\$}(1)\right]-\eta_1}-K\right),0\right\}\right] \\ &= \mathbb{E}_0\left[e^{-y_0^{\$}(1)-\frac{1}{2}\lambda_0' R\lambda_0-\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\eta_1}\max\left\{\left(e^{-\mathbb{E}_0\left[y_1^{\$}(1)\right]-\eta_1}-K\right),0\right\}\right]e^{\frac{1}{2}\left(\lambda_0' R\lambda_0-\frac{\left(\sigma' R\lambda_0\right)^2}{\sigma' R\sigma}\right)}\end{aligned}$$

The option will be exercised if and only if the following holds

$$\eta_1 < -\log\left(K\right)-\mathbb{E}_0\left[y_1^{\$}\left(1\right)\right],$$

which occurs with probability

$$\Phi\left(\frac{-\log\left(K\right)-\mathbb{E}_0\left[y_1^{\$}\left(1\right)\right]}{\sqrt{\sigma' R\sigma}}\right)\equiv\Phi\left(x^{\star}\right).$$

We proceed:

$$\begin{aligned}C_0\left(K\right) &= e^{\frac{1}{2}\left(\lambda_0' R\lambda_0-\frac{\left(\sigma' R\lambda_0\right)^2}{\sigma' R\sigma}\right)}\mathbb{E}_0\left[e^{-y_0^{\$}(1)-\frac{1}{2}\lambda_0' R\lambda_0-\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\eta_1}\left(e^{-\mathbb{E}_0\left[y_1^{\$}(1)\right]-\eta_1}-K\right)I_{\left(\eta_1/\sqrt{\sigma' R\sigma}<x^{\star}\right)}\right] \\ &= e^{\frac{1}{2}\left(\lambda_0' R\lambda_0-\frac{\left(\sigma' R\lambda_0\right)^2}{\sigma' R\sigma}\right)}\int_{-\infty}^{x^{\star}}e^{-y_0^{\$}(1)-\mathbb{E}_0\left[y_1^{\$}(1)\right]-\frac{1}{2}\lambda_0' R\lambda_0-\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}x-\sqrt{\sigma' R\sigma}x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x \\ &\quad -e^{\frac{1}{2}\left(\lambda_0' R\lambda_0-\frac{\left(\sigma' R\lambda_0\right)^2}{\sigma' R\sigma}\right)}\int_{-\infty}^{x^{\star}}Ke^{-y_0^{\$}(1)-\frac{1}{2}\lambda_0' R\lambda_0-\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\mathrm{d}x,\end{aligned}$$

where we use that $\eta_{1/}\sqrt{\sigma' R\sigma}$ is standard normally distributed. Rewriting and using that:

$$-2y_0^{\$}\left(2\right) = -y_0^{\$}\left(1\right)-\mathbb{E}_0\left[y_1^{\$}\left(1\right)\right]+\frac{1}{2}\sigma' R\sigma+\sigma' R\lambda_0,$$

we obtain:

$$\begin{aligned}C_0\left(K\right) &= P_0^{\$}\left(2\right)\int_{-\infty}^{x^{\star}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(x+\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}+\sqrt{\sigma' R\sigma}\right)^2}\mathrm{d}x-KP_0^{\$}\left(1\right)\int_{-\infty}^{x^{\star}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(x+\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}\right)^2}\mathrm{d}x \\ &= P_0^{\$}\left(2\right)\Phi\left(x^{\star}+\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}+\sqrt{\sigma' R\sigma}\right)-KP_0^{\$}\left(1\right)\Phi\left(x^{\star}+\frac{\sigma' R\lambda_0}{\sigma' R\sigma}\sqrt{\sigma' R\sigma}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Using the definition of $x^\star$, we conclude that the option value is given by:

$$\begin{aligned} C_0(K) &= P_0^{\$}(2)\,\Phi(d_1) - K P_0^{\$}(1)\,\Phi(d_2), \\ d_1 &\equiv \frac{-\log(K) - \mathbb{E}_0\left[y_1^{\$}(1)\right] + \sigma' R\sigma + \sigma' R\lambda_0}{\sqrt{\sigma' R\sigma}}, \\ &= \frac{\log\left(P_0^{\$}(2)/K\right) + y_0^{\$}(1) + \frac{1}{2}\sigma' R\sigma}{\sqrt{\sigma' R\sigma}}, \\ d_2 &\equiv d_1 - \sqrt{\sigma' R\sigma}, \end{aligned}$$

where the second line for d_1 uses the pricing formula of a two-period bond. Now using $K = 1/\left(1+q_0^{FRMP}\right)$ and the fact that the investor has $B\left(1+q_0^{FRMP}\right)$ of these options, yields the value of the prepayment option:

$$B\mathbb{E}_0\left[M_1^{\$}M_2^{\$}\max\left\{\left(q_0^{FRMP} - q_1^{ARM}\right),0\right\}\right] = B\left(1+q_0^{FRMP}\right)C_0\left(1/\left(1+q_0^{FRMP}\right)\right). \tag{4.22}$$

4.D Multi-Period Model

In this appendix, we consider a more realistic, multi-period extension of the simple model in Section 4.3. It has power utility preferences and features an exogenous moving probability. We use this model (i) to study the role of the time discount factor and the moving rate, and (ii) to solve for the relationship between the cross-sectional distribution over risk aversion parameters and the aggregate ARM share.

4.D.1 Setup

The household problem Household j chooses the mortgage contract, $i \in \{FRM, ARM\}$, to maximize expected lifetime utility over real consumption:

$$U_j^i = E_0 \sum_{t=1}^{T} \beta^t (1-\xi)^{t-1} \frac{\left(C_t^i\right)^{1-\gamma_j}}{1-\gamma_j}, \tag{4.23}$$

$$C_t^i = L - q_t^i/\Pi_t, \text{ for } t \in \{1,\ldots,T-1\} \tag{4.24}$$

$$C_T^i = 1 + L - \left(1+q_T^i\right)/\Pi_T \tag{4.25}$$

where β is the (monthly) subjective discount rate, ξ is the (monthly) exogenous moving rate, and γ_j is the coefficient of relative risk aversion. We consider constant real labor income L. We normalize the nominal outstanding balance to one, which makes q_t^i both the nominal mortgage rate and nominal mortgage payment at time t for contract i. This setup incorporates utility up until a move. The certainty-equivalent consumption, $\tilde{C}^i$, is given by:

$$\tilde{C}_j^i = \left(U_j^i \Big/ \sum_{t=1}^{T} \frac{\beta^t (1-\xi)^{t-1}}{1-\gamma_j}\right)^{1/(1-\gamma_j)}. \tag{4.26}$$

We are interested in the certainty-equivalent consumption differential $\tilde{C}_j^{ARM} - \tilde{C}_j^{FRM}$.

Bond Pricing Following Koijen, Nijman, and Werker (2007a), we consider a continuous-time, two-factor essentially affine term structure model. The factors $X_t = [Z_{1t}, Z_{2t}]'$ are identified with the real rate

and expected inflation, respectively. The model can be discretized exactly to a VAR(1)-model:

$$Z_t = \mu + \Phi Z_{t-1} + \Sigma\varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0, I_{3\times 3}), \tag{4.27}$$

where the third element of the state is realized inflation, $Z_{3t} = \log \Pi_t - \log \Pi_{t-1}$. The $\tau-$month bond price at time t is exponentially affine in X_t:

$$P_t^{\$}(\tau) = \exp\{A_\tau + B_\tau' X_t\}, \tag{4.28}$$

where $A_\tau = A(\tau/12)$ and $B_\tau = B(\tau/12)$, with $A(\cdot)$ and $B(\cdot)$ derived in Appendix A of Koijen, Nijman, and Werker (2007a).

Mortgage Pricing At time t the lender of the FRM receives

$$q^{FRM}(1-\xi)^{t-1} + (1-\xi)^{t-1}\xi, \tag{4.29}$$

where $(1-\xi)^{t-1}$ is the probability that loan has not been prepaid before time t and $(1-\xi)^{t-1}\xi$ is the probability it is prepaid at time t. Imposing a zero-profit condition, a mortgage contract of T periods has the following FRM rate:

$$q^{FRM} = \frac{1 - \sum_{t=1}^{T-1}(1-\xi)^{t-1}\xi P_0^{\$}(t) - (1-\xi)^{T-1} P_0^{\$}(T)}{\sum_{t=1}^{T-1}(1-\xi)^{t-1} P_0^{\$}(t) + (1-\xi)^{T-1} P_0^{\$}(T)}. \tag{4.30}$$

For the monthly ARM rate we have $q_t^{ARM} = P_t^{\$}(1)^{-1} - 1$.

4.D.2 Calibration

The term structure parameters are taken from Koijen, Nijman, and Werker (2007a). As is the case for the VAR estimates in the main text, the correlation between the yield spread and the bond risk premium is low in the model (-7%). Real labor income, L, is held constant at 0.42. To obtain a theoretically well-defined problem we assume a minimum subsistence consumption level of 0.05/12 per month. The exogenous monthly moving probability is is set at 1% per month ($(1-\xi)^{12} - 1 = 11.36\%$ per year). We consider different values for the coefficient of relative risk aversion, γ, and the monthly subjective discount factor, β.

4.D.3 Effect of the Subjective Discount Factor and Moving Rates

We generate $N = 1000$ starting values for the state vector at time zero, Z_0, by simulating forward $M = 60$ months from the unconditional mean for the state vector (0_{4*1}) for each of the N paths. Next, we compute the expected utility differential of the ARM and FRM contracts. Expected utilities are computed by averaging realized utilities in $K = 100$ simulated paths (where the same shocks apply to all $N = 1000$ starting values).

Figure 4.E plots the R^2 of regressing the model's certainty-equivalent consumption differential between the ARM and FRM contracts on the model's bond risk premium (solid line) or on the model's yield spread (dashed line). Each point corresponds to a different value of the annualized subjective time discount factor β^{12}, between 0.5 and 1. The coefficient of relative risk aversion is set at $\gamma = 5$. For low values of the subjective discount factor ($\beta < .70$), the slope of the yield curve has a stronger relationship to the relative desirability of the ARM. However, for more realistic and more conventional values of the subjective discount factor, say between 0.9 and 1.0, the bond risk premium is the key determinant of mortgage choice. We

have also experimented with an upward sloping labor income profile, as in Cocco, Gomes, and Maenhout (2005), and found a similar cut-off rule. A similar result holds when we vary the moving rate instead of the subjective time discount factor: below 10% per month, the risk premium is the more important predictor. For empirically relevant moving rates below 2%, the risk premium is the only relevant predictor.

4.D.4 Heterogeneous Risk Aversion Level

For each month in our sample period we determine the level of risk aversion that makes an investor indifferent between the ARM and the FRM. Starting values for the vector of state variables, Z, are from Koijen, Nijman, and Werker (2007a). The utility differential of an ARM and an FRM is computed as described above. The monthly subjective discount factor is set at $\beta = 0.96^{1/12} \approx 0.9966$. We assume a log-normal cross-sectional distribution for the risk aversion level:

$$\log(\gamma) \sim \mathcal{N}\left(\mu_\gamma, \sigma_\gamma^2\right), \tag{4.31}$$

which implies that our model predicts the following ARM share:

$$ARM_t^{pred}\left(\log\left(\gamma_t^*\right); \mu_\gamma, \sigma_\gamma\right) = \Phi\left(\frac{\log\left(\gamma_t^*\right) - \mu_\gamma}{\sigma_\gamma}\right) \tag{4.32}$$

where Φ is the standard normal cumulative density function and where households with a risk aversion smaller than the cutoff γ_t^* choose the ARM. More conservative households choose the FRM.

We determine μ_γ and σ_γ by minimizing the squared prediction error over the sample period (1985:1-2005:12) and estimate a location parameter $\hat{\mu}_\gamma = 5.0$ and a scale parameter $\hat{\sigma}_\gamma = 2.9$. The median level of risk aversion implied by this distribution equals $\exp\left(\hat{\mu}_\gamma\right) = 155$. Interestingly, regressing the actual ARM share on the predicted ARM share yields a constant and slope coefficient of 0.03 and 0.90 respectively, which are not significantly different from theoretical implied values of 0 and 1 respectively.

The cutoff log risk aversion level has a sample mean of $\mu_{\gamma^*} = 3.37$ and a sample standard deviation of $\sigma_{\gamma^*} = 0.73$. The predicted increase in the ARM share from a one standard deviation increase in the log indifference risk aversion level around its mean is given by:

$$ARM^{pred}\left(\mu_{\gamma^*} + 0.5\sigma_{\gamma^*}; \hat{\mu}_\gamma, \hat{\sigma}_\gamma\right) - ARM^{pred}\left(\mu_* - 0.5\sigma_{\gamma^*}; \hat{\mu}_\gamma, \hat{\sigma}_\gamma\right) = 8.6\% \tag{4.33}$$

This 8.6% is very close to the slope coefficient we reported in Table 4.2, Rows 3-6. In conclusion, the model can explain the observed average 28% ARM share and the observed sensitivity of the ARM share to the bond risk premium with a mean log risk aversion of 5 and a standard deviation of log risk aversion of 2.9.

We conjecture that these values would be lower in a model where labor income risk were negatively correlated with the real rate. In that case, the ARM would be more risky because ARM payments would be high when labor income is low. A lower risk aversion would be needed to choose the FRM. Put differently, the (relatively low) observed ARM share could be justified with a lower mean risk aversion.

4.E Tables and figures

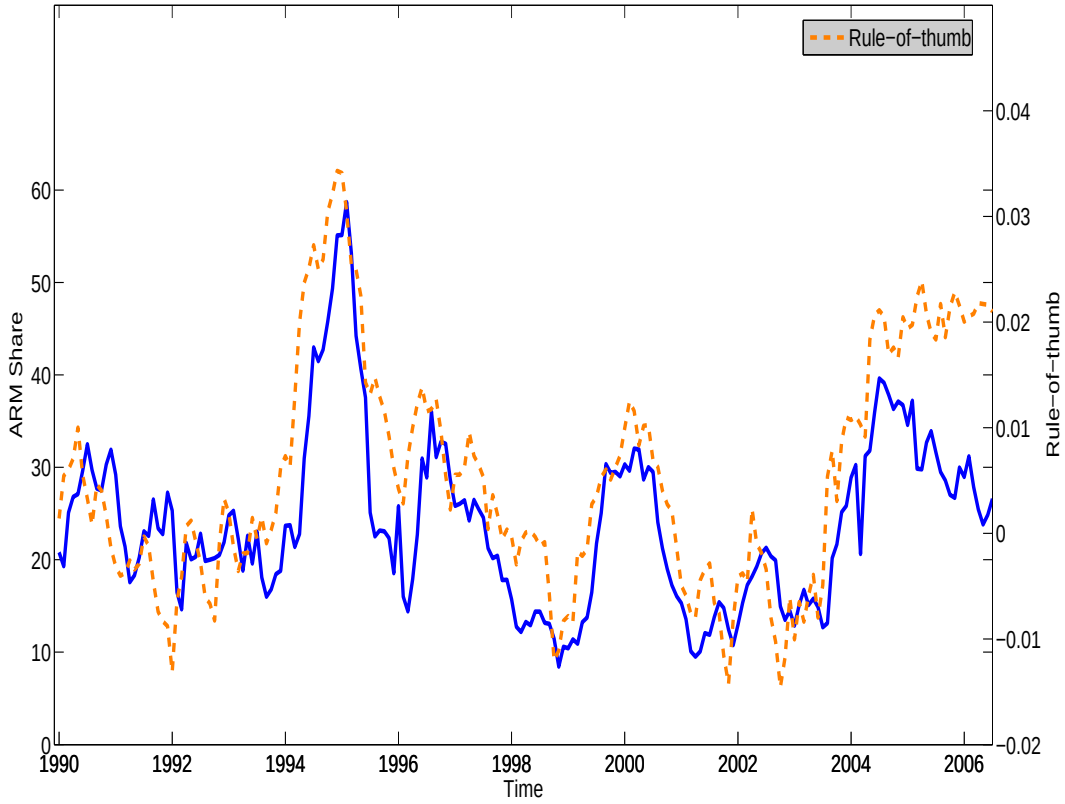


Figure 4.1: Household decision rule and the ARM share
The solid line corresponds to the ARM share in the US, and its values are depicted on the left axis. The dashed line displays the household decision rule $\kappa_t(3,5)$. It is computed as the difference between the 5-year Treasury yield and the 3-year moving average of the 1-year Treasury yield. The time series is monthly from 1989.12 to 2006.6.

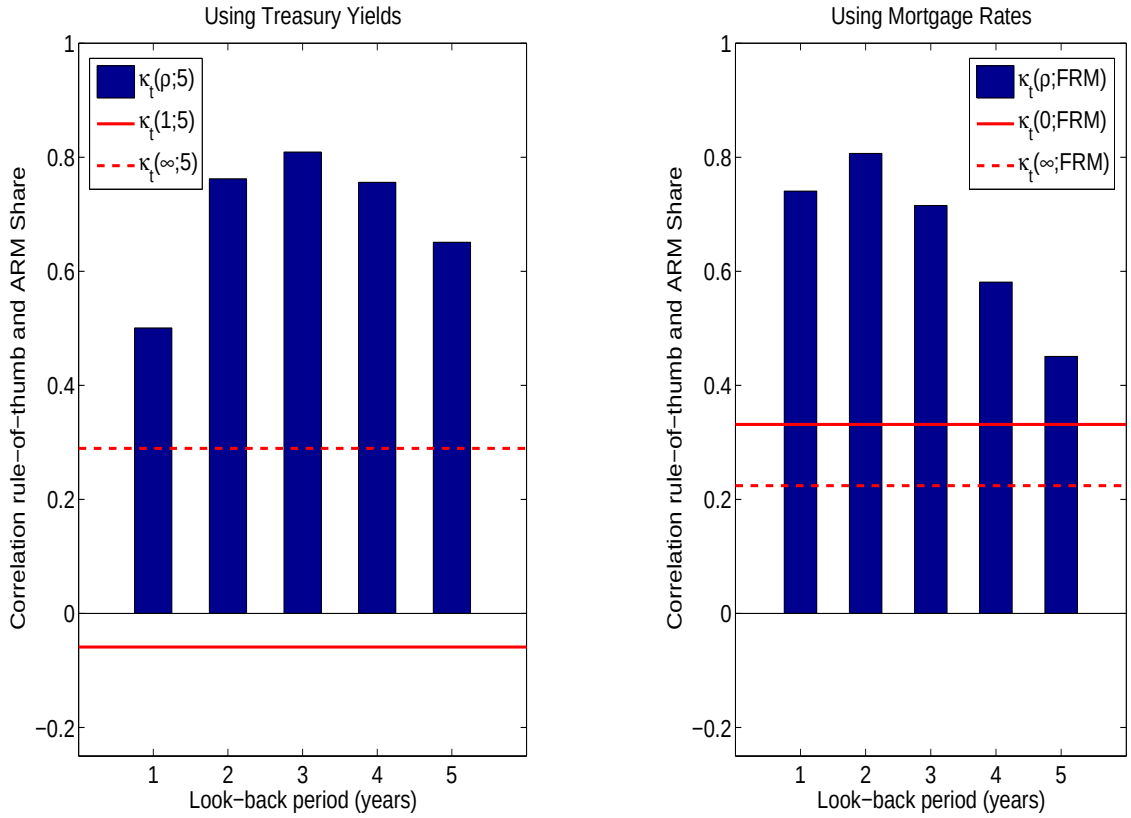


Figure 4.2: Correlation of the household decision rule and the ARM share for different look-back horizons ρ

The figure plots the correlation of the household decision rule $\kappa_t(\rho;\tau)$ with the ARM share. The blue bars correspond to $\rho = 1$, 2, 3, 4, and 5 years. The red line corresponds to the correlation between the 5-1 year yield spread (i.e., $\tau = 5$ and $\rho = 1$) and the ARM share. The red dashed line depicts the correlation between the 5-year yield and the ARM share (i.e., $\tau = 5$ and $\rho = \infty$). The left panel uses Treasury yields as yield variable ($\tau = 5$), while the right panel uses the effective 1-year ARM and effective 30-year FRM rates ($\tau = FRM$). The results are shown for the period 1989.12-2006.6, the longest sample for which all measures are available.

$\kappa_t(3;5)$	$y^{\$}(5) - y^{\$}(1)$	$y^{\$}(5)$	BAL	FICO	LTV	Regional dummies	% correctly classified
0.43 [253]						No	69.4
			-0.05 [21]	-0.05 [28]	0.17 [100]	Yes	61.7
0.42 [244]			-0.01 [4]	-0.08 [45]	0.13 [72]	Yes	68.8
	0.06 [38]					No	59.8
	0.09 [53]		-0.05 [23]	-0.06 [30]	0.19 [106]	Yes	62.1
0.65 [299]	0.43 [206]		-0.00 [2]	-0.11 [58]	0.17 [90]	Yes	70.9
		-0.30 [171]				No	64.7
		-0.33 [179]	-0.05 [22]	-0.09 [46]	0.20 [110]	Yes	66.6
0.54 [290]		-0.47 [237]	-0.00 [1]	-0.15 [71]	0.16 [80]	Yes	71.6

Table 4.1: Probit regressions of the ARM share in loan-level data
This table reports slope coefficients, robust t-statistics (in brackets), and R^2 statistics for probit regressions of an ARM dummy on a constant and one or more regressors, reported in the first column. The regressors are $\kappa(3,5)$, the household decision rule formed with a 5-year Treasury yield and a 3-year average of past 1-year Treasury yield data, the loan balance at origination (BAL), the loan's credit score at origination (FICO), the loan's loan-to-value ratio (LTV), the long-term interest rate (5-year Treasury yield), and the 5-1 year Treasury yield spread. The seventh column indicates when we include four regional dummies for the biggest mortgage markets (California, Florida, New York, and Texas). All independent variables have been normalized by their standard deviation. The sample consists of 654,368 mortgage loans originated between 1994-2006.6.

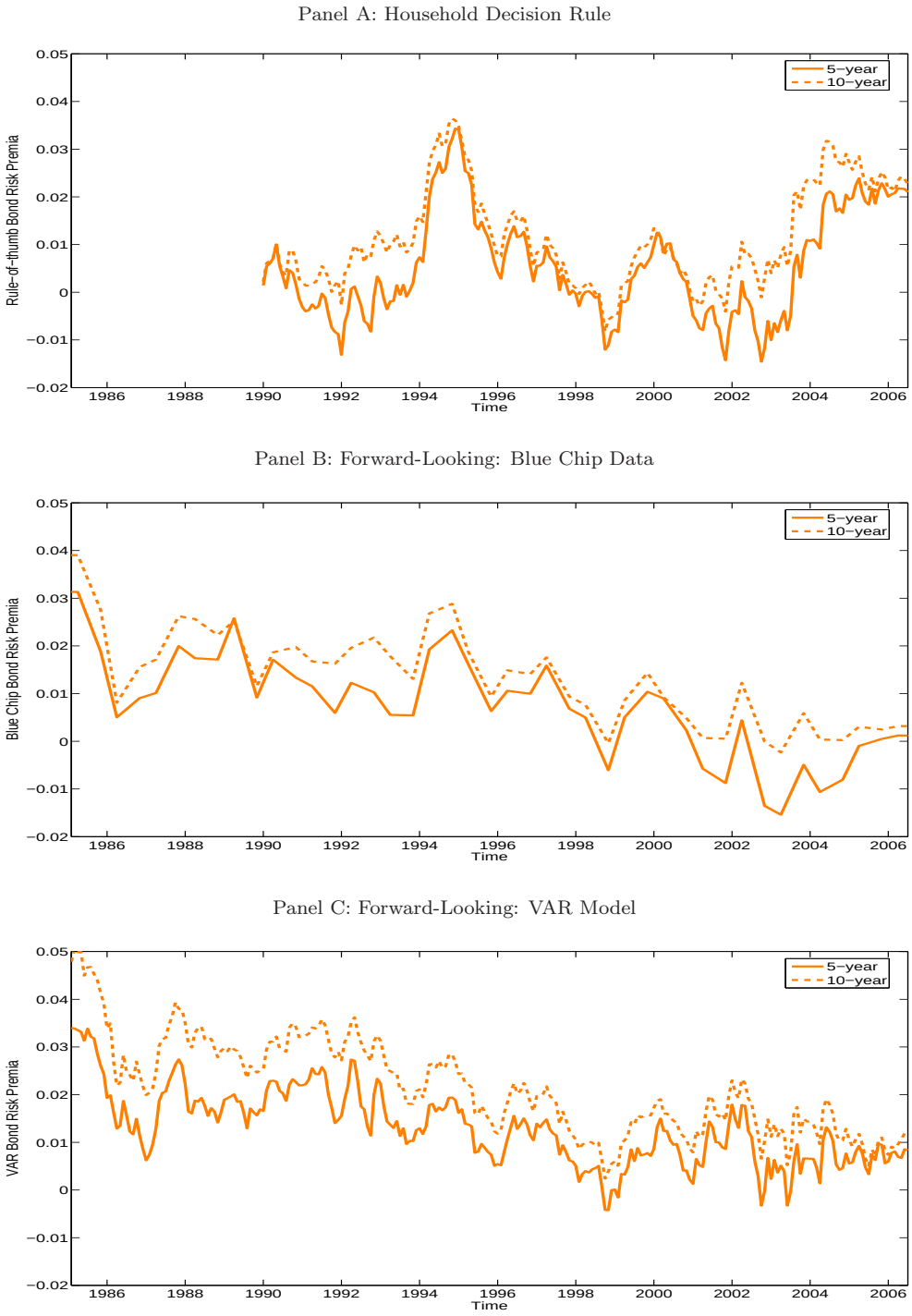


Figure 4.3: Three measures of the nominal bond risk premium
Each panel plots the 5-year and the 10-year nominal bond risk premium. The average expected future nominal short rates that go into this calculation differ in each panel. In the top panel we use adaptive expectations with a three-year look-back period. In the middle panel we use Blue Chip forecasters data. In the bottom panel we use forecasts formed from a VAR model.

			slope	t-stat	R^2
1.	Househ. Decis. Rule	$\kappa_t(3,5)$	7.88	7.08	71.23
2.		$\kappa_t(3,10)$	7.70	7.47	68.03
3.	Blue Chip	$\phi_t^{\$}(5)$	8.63	3.91	40.25
4.		$\phi_t^{\$}(10)$	8.89	4.22	42.62
5.	VAR	$\phi_t^{\$}(5)$	7.73	4.16	32.21
6.		$\phi_t^{\$}(10)$	8.07	3.91	35.13
7.	Slope	$y_t^{\$}(5) - y_t^{\$}(1)$	0.46	0.21	0.11
8.		$y_t^{\$}(10) - y_t^{\$}(1)$	-0.66	-0.32	0.23
9.	Long yield	$y_t^{\$}(5)$	8.37	3.76	37.76
10.		$y_t^{\$}(10)$	8.53	3.85	39.26
11.	Mortgage rates	$\kappa_t(2,FRM)$	7.26	9.37	60.40
12.		$\kappa_t(3,FRM)$	6.28	4.99	45.28
13.		$y_t^{\$}(FRM) - y_t^{\$}(ARM)$	8.09	3.17	35.31
14.		$y_t^{\$}(FRM) - y_t^{\$}(ARM)$ orth.	8.75	3.86	41.28
15.		$y_t^{\$}(FRM)$	7.81	3.71	32.87
16.	Other Rules-of-Thumb	FRM rule	6.00	3.74	22.54
17.		ARM rule	3.13	2.42	6.12

Table 4.2: The ARM share and the nominal bond risk premium
This table reports slope coefficients, Newey-West t-statistics (12 lags), and R^2 statistics for regressions of the ARM share on a constant and the regressors reported in the first column. The regressors are the τ-year nominal bond risk premium $\phi_t^{\$}(\tau)$, measured in three different ways. We consider $\tau = 5$ and $\tau = 10$ years. The first measure is based on the household decision rule with a 3-year look-back period (rows 1-2). The second measure is based on Blue Chip forecast data (rows 3 and 4) and the third measure is based on the VAR (rows 5-6). Rows 7 and 8 show regressions of the ARM share on the τ-1-year yield spread $y_t^{\$}(\tau) - y_t^{\$}(12)$. Rows 9 and 10 use the τ-year nominal yield, $y_t^{\$}(\tau)$, as predictor. Rows 11 and 12 use the household decision rule computed using the effective 30-year FRM rate and the effective 1-year ARM rate, with a look-back period of 2 years in Row 11 and three years in Row 12. Row 13 uses the difference between the FRM rate $y_t^{\$}(FRM)$ and the ARM rate $y_t^{\$}(ARM)$, while row 15 uses $y_t^{\$}(FRM)$ as independent variable. Row 14 uses the component of the FRM-ARM spread that is orthogonal to the 10-1 Treasury bond spread. Rows 16 and 17 consider two other rules-of-thumb. The FRM rule takes the current FRM rate minus the three-year moving average of the FRM rate (row 16). The ARM rule in Row 17 does the same for the ARM rate. In all rows, the regressor is lagged by one period, relative to the ARM share. All independent variables have been normalized by their standard deviation. The sample is 1985.1-2006.6, except for rows 1 and 2 and 11 and 12, where we use 1989.12-2006.6, the sample for which the household decision rules are available.

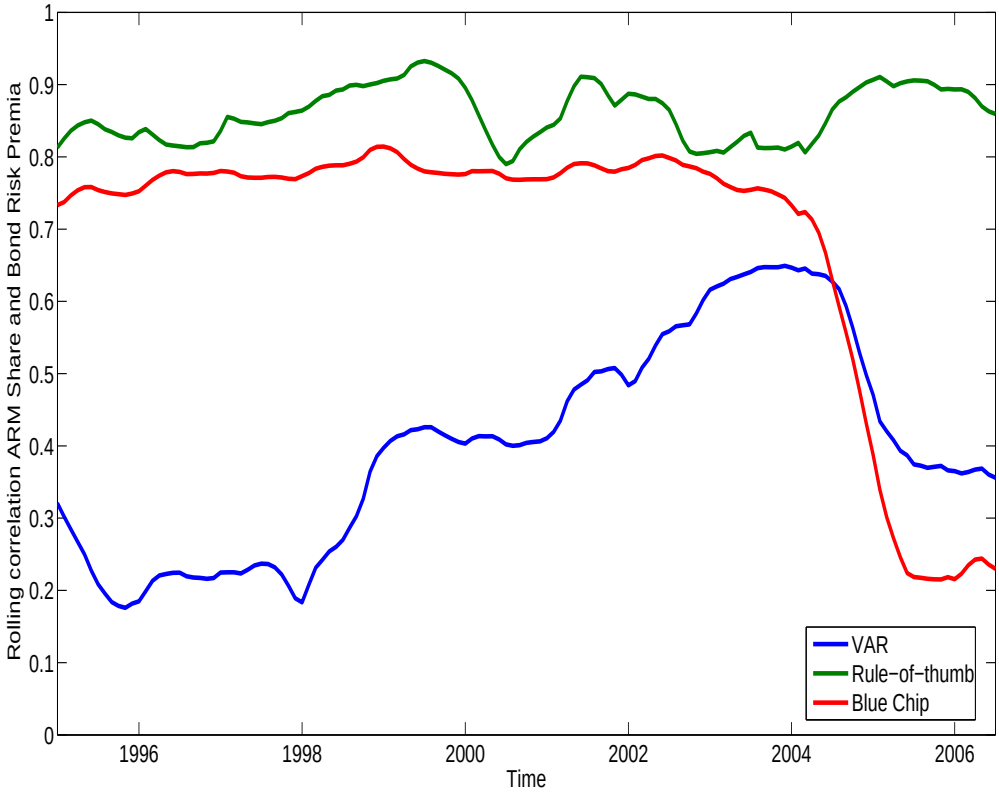


Figure 4.4: Rolling window correlations
The figure plots 10-year rolling window correlations of each of the three bond risk premium measures with the ARM share. The top line is for the household decision rule (dotted), the middle line is for the measure based on Blue Chip forecasters data (solid), and the bottom line is based on the VAR (dashed). The first window is based on the 1985-1995 data sample.

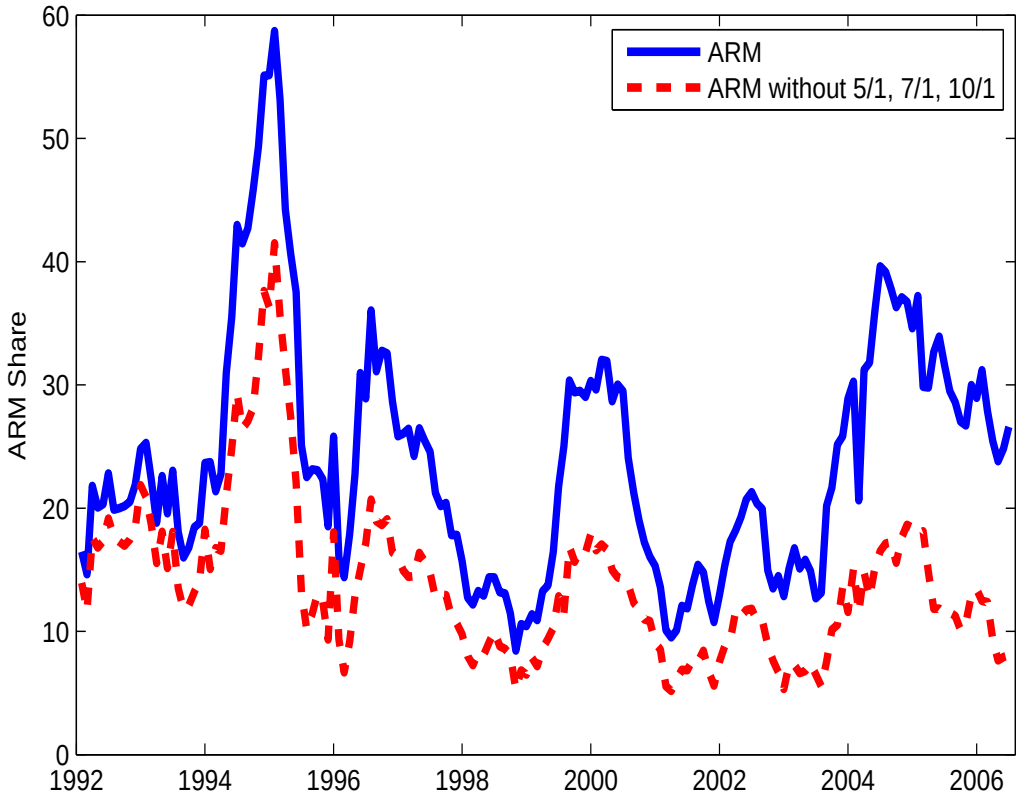


Figure 4.5: Product innovation in the mortgage market
The solid line plots our benchmark ARM share, which includes all hybrid mortgage contracts, between 1992.1 and 2006.6. The dashed line excludes all hybrids with an initial fixed-rate period of more than three years. The data are from the Monthly Interest Rate Survey compiled by the Federal Housing Financing Board.

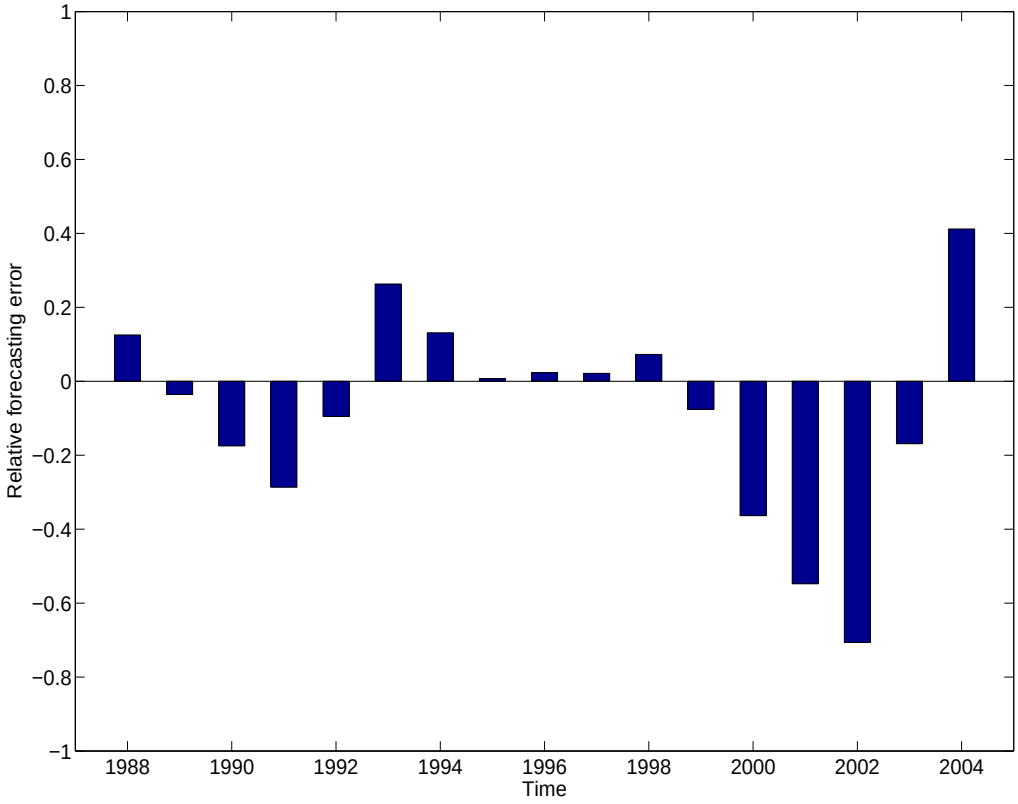


Figure 4.6: Errors in predicting future real rates
The figure plots forecast errors in expected future real short rates. The forecast error is computed using Blue Chip forecast data. The average expected future real short rate is calculated as the difference between the Blue Chip consensus average expected future nominal short rate and the Blue Chip consensus average expected future inflation rate. The realized real rate is computed as the difference between the realized nominal rate and the realized expected inflation, which are measured as the one-quarter ahead inflation forecast. The realized average future real short rates are calculated from the realized real rates. The forecast errors are scaled by the nominal short rate to obtain relative forecasting errors. The forecast errors are based on two-year ahead forecasts.

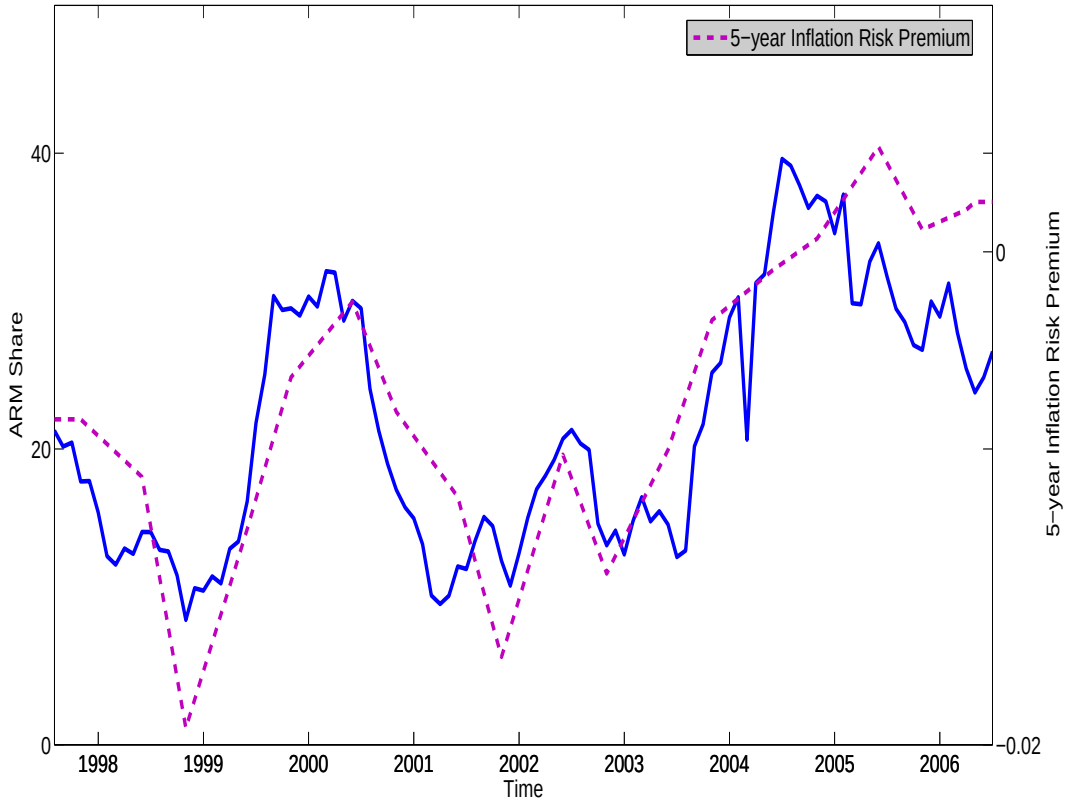


Figure 4.7: The inflation risk premium and the ARM share
The figure plots the fraction of all mortgages that are of the adjustable-rate type against the left axis (solid line), and the inflation risk premium (dashed line) against the right axis. The inflation risk premium is computed as the difference between the 5-year nominal bond yield, the 5-year real bond yield and the expected inflation. The real 5-year bond yield data are from McCulloch and start in January 1997. The inflation expectation is the Blue Chip consensus average future inflation rate over the next 5 years.

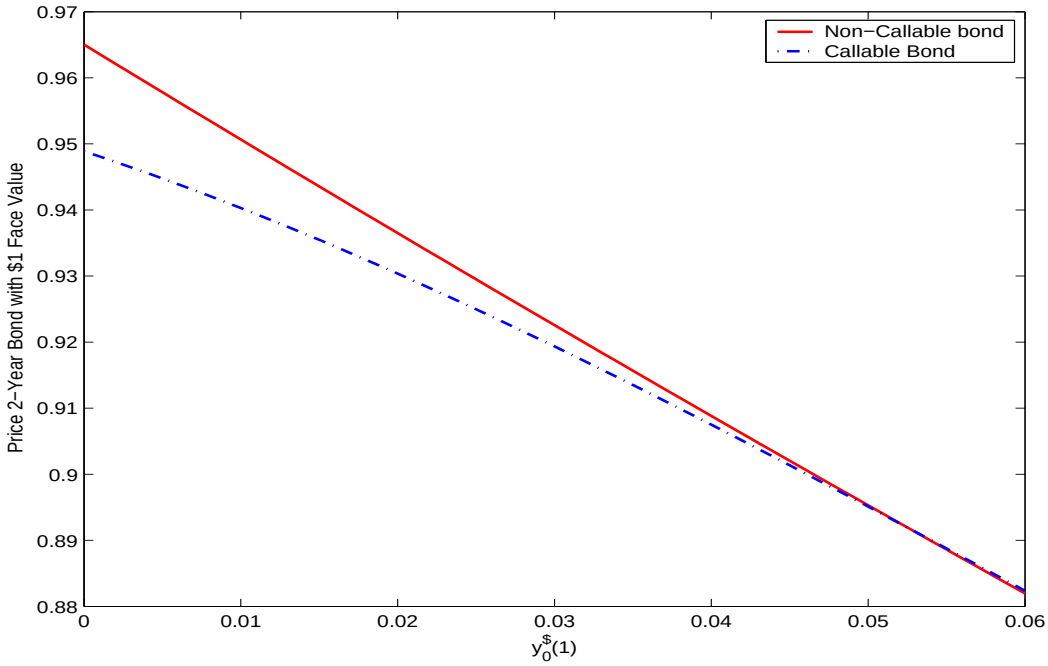


Figure 4.8: Price sensitivity to changes in the nominal interest rate
The figure plots the price sensitivities of the FRM contract with and without prepayment to the nominal interest rate, $y_0^{\$}(1)$. The mortgage values are determined within the model of Section 4.6.1. The analogous fixed-income securities are a regular bond (FRM without prepayment) and a callable bond (FRM with prepayment).

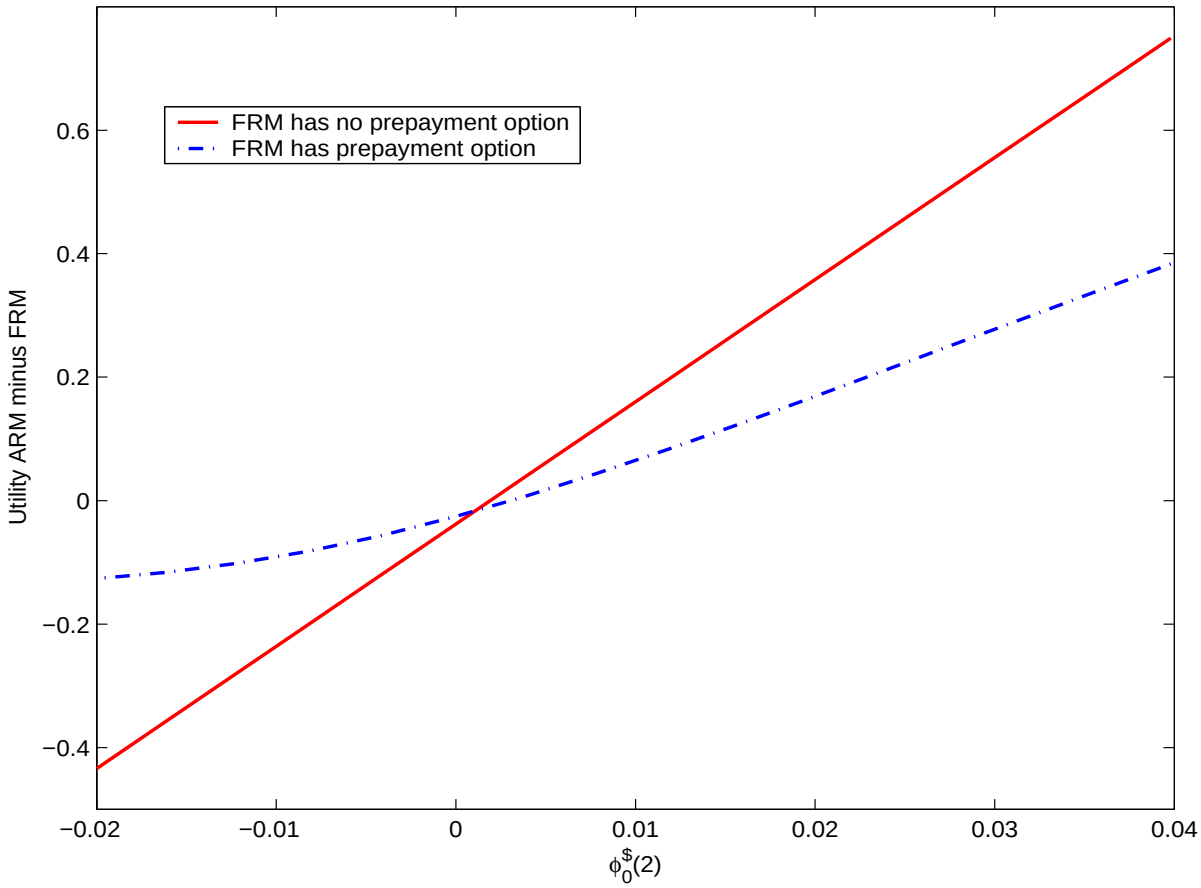


Figure 4.9: Utility difference between ARM and FRM - prepayment
The figure plots the utility difference between an ARM contract and an FRM contract without prepayment as well as the utility difference between an ARM contract and an FRM contract with prepayment.

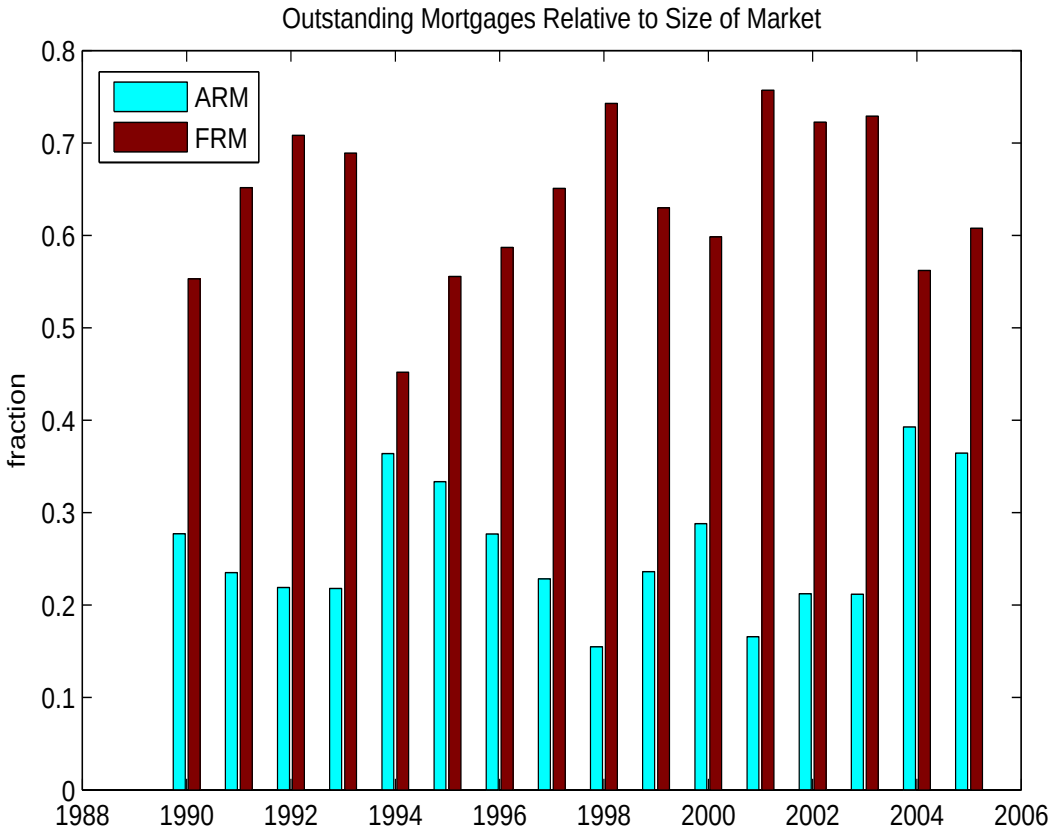


Figure 4.10: Mortgage originations in the US
The figure plots the volume of conventional ARM and FRM mortgage originations in the US between 1990 and 2005, scaled by the overall size of the mortgage market. Data are from the Office of Federal Housing Finance Enterprise Oversight (OFHEO).

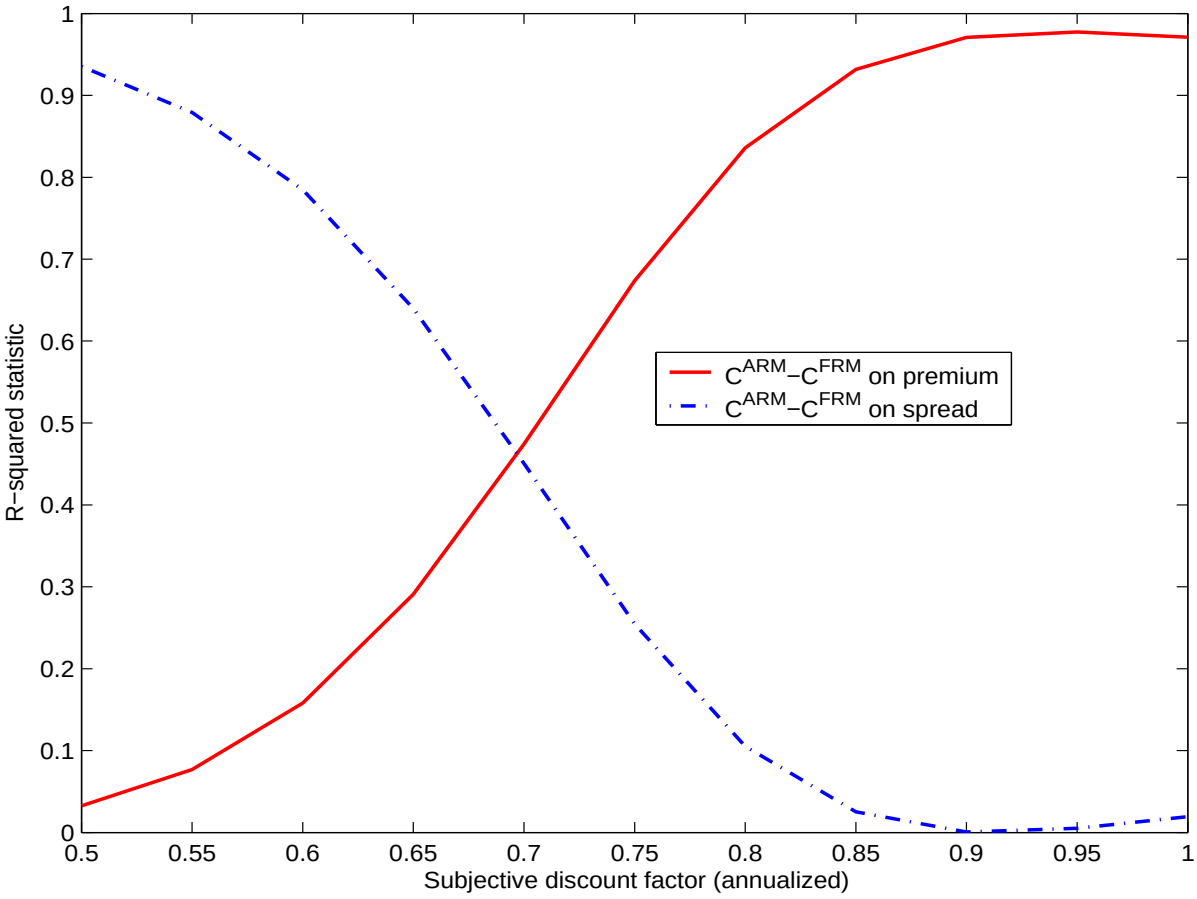


Figure 4.11: Effect of the rate of time preference
Each point in the figure corresponds to the R^2 of a regression of the certainty equivalent consumption difference between an ARM contract and an FRM contract on either the bond risk premium (solid line) or one the yield spread (dashed line). The annualized subjective discount factor β^{12}, on the horizontal axis, is varied between 0.5 and 1. The time series are generated from a model, which is a multi-period extension of the model in Section 4.3. The coefficient of relative risk aversion is $\gamma = 5$. The exogenous moving probability is held constant at 1% per month.

Chapter 5

Predictive Regressions: A Present-Value Approach

Abstract

We use a closed-form present-value model to estimate the time series of expected returns and expected dividend growth rates. By imposing the economic restrictions implied by this structural model, we show that the aggregate price-dividend ratio has strong predictive power for both future stock market returns and future dividend growth, with R-squared values of 18% and 16%, respectively. In contrast, reduced-form, statistical models of returns, dividend growth rates, and the price-dividend ratio generally result in weak predictive power for future returns and no predictive power for future dividend growth. We find that expected dividend growth has both a transient and a persistent component. Our procedure allows us to bypass the standard Vector Auto Regressions in which the instruments can be misspecified and/or suffer from a selection bias. We then decompose each historical stock return into shocks to expected returns, unexpected dividend shocks, and shocks to both components of expected dividend growth, which reveals the economic drivers of asset prices.

5.1 Introduction

Conventional wisdom states that the price-dividend ratio has no predictive power for future dividend growth rates and only weak predictive power for future returns. This conclusion is generally based on reduced-form Vector Auto Regression (VAR) models of returns, dividend growth, and the price-dividend ratio. We, instead, use a new class of closed-form present-value models together with simple non-linear filtering techniques to estimate the time series of expected returns and expected dividend growth rates. We are the first to show that by imposing the economic restrictions that follow from a structural present-value model, the price-dividend ratio predicts both future annual stock returns and future annual dividend growth rates, with R-squared values of 18% and 16%, respectively.[1] We find that expected

[1]Our main results are based on nominal aggregate dividends as in Boudoukh, Michaely, Richardson, and Roberts (2004) to account for share repurchases, but we obtain similar results when using real data and

dividend growth has both a transient and a persistent component. Additionally, we show that these cross-equation restrictions help to alleviate many of the statistical issues that have been central in the predictability literature. We then use the present-value model to decompose historical stock returns into shocks to expected returns, unexpected dividend shocks, and shocks to both components of expected dividend growth. This decomposition reveals the economic drivers of asset prices.

The main mechanism underlying this result is as follows. Variation in the price-dividend ratio stems from expected return variation and fluctuations in expected dividend growth rates. In our model, the price-dividend ratio (PD_t) is linear in expected returns ($\hat{\mu}_t$) and expected dividend growth rates ($\hat{g}_t$):

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t, \tag{5.1}$$

where A, B_1, and B_2 depend on the structural parameters describing the dynamics of expected returns and dividend growth rates. Equation (5.1) shows that if we can obtain an estimate of $\hat{\mu}_t$ and we combine this with the observed value of the price-dividend ratio, the present-value model delivers an estimate of expected growth rates: $\hat{g}_t = B_2^{-1}\left(PD_t - A - B_1\hat{\mu}_t\right)$. For instance, we can use the full history of returns and the price-dividend ratio to form an estimate of expected returns. The present-value restriction then provides an estimate of expected growth rates. Likewise, given an estimate for expected growth rates and the current price-dividend ratio, we obtain an estimate of expected returns: $\hat{\mu}_t = B_1^{-1}\left(PD_t - A - B_2\hat{g}_t\right)$. In this case, we can use the full history of dividend growth rates and the price-dividend ratio to estimate $\hat{g}_t$, which results in an estimate of $\hat{\mu}_t$ by imposing the present-value restriction. In our estimation we combine these two procedures. As such we develop a method that optimally combines past returns, past growth rates, and the current price-dividend ratio to form estimates of both expected returns and expected growth rates that satisfy the present-value restriction. These estimates outperform estimates based on the price-dividend ratio alone. Our main insight is that better predictors of returns, once combined with a structural present-value model, improve estimates of expected growth rates, and vice versa. Empirically, we show that this implies that both returns and dividend growth rates contain an economically significant predictable component. The information set can be extended with additional instruments if one has prior views on which instruments to include. As before, any instrument that improves the prediction of future returns directly improves the estimate of dividend growth rates, and the opposite is true for instruments predicting dividend growth rates.

A second important contribution is that the estimates of the present-value model are

when share repurchases are not taken into account.

accurate and virtually unbiased. This is in sharp contrast to standard predictive regressions that are known to be biased (Stambaugh, 1999) and grossly inefficient. The mere fact that the present-value relationship can sharpen estimates is not new to this paper (Cochrane (2006) and Campbell and Yogo (2006)). However, these results have generally been established in models in which either expected dividend growth rates are assumed to constant or the persistence of expected returns and expected growth rates are assumed to be equal. We do not impose either of these restrictions. Further, we are the first to show that by imposing the present-value relationship the bias in the predictive coefficient reduces substantially.

In the 1980s, the price-dividend ratio was found to predict future stock returns (e.g., Fama and French (1988)) spurring a large literature on whether or not returns are predictable.[2] Proponents of return predictability argue that, given the large variation of the price-dividend ratio, returns will have to be predictable because dividend growth rates are not (Cochrane (2006)). Regressing dividend growth rates on the lagged price-dividend ratio even leads to the incorrect sign for the predictive coefficient. That is, a high price-dividend ratio seems to predict low as opposed to high future dividend growth. We show that both return predictability and dividend growth predictability are perfectly consistent with the observed dynamics for the price-dividend ratio. In fact, our model provides an explanation for the perverse sign in the dividend growth regression. We find that expected returns and expected growth rates are both strongly time-varying and the two series are highly positively correlated.[3] Even though changes in expected returns and changes in expected dividend growth rates affect the price-dividend ratio with opposite (theoretically correct) signs, they largely offset each other because they have such a high correlation. Because expected returns are somewhat more persistent, it is usually the change in expected returns that dominates. As a consequence, regressing dividend growth rates on the lagged price-dividend ratio will lead to the incorrect sign of the predictive coefficient. We can conclude that predictive regressions suffer from an errors-in-variables problem.[4] This errors-in-variables problem is most severe when we try to predict dividend growth rates, but it also affects the predictive regression for returns.[5] We, instead, use a present-value model combined with non-linear filtering techniques to recover both time series and show that this results in strong predictors of both future returns and future dividend growth rates.[6]

[2]Important recent contributions to this debate include Goyal and Welch (2003), Goyal and Welch (2006), Cochrane (2006), Campbell and Thompson (2007), Campbell and Yogo (2006), and Lettau and van Nieuwerburgh (2006).

[3]Lettau and Ludvigson (2005) reach a similar conclusion using CAY and CDY to predict future returns and future dividend growth rates.

[4]This was also pointed out by Goetzman and Jorion (1995).

[5]It is important to note that the errors-in-variables problem does not disappear as the sample size increases. Our approach both resolves this errors-in-variables problem and additionally alleviates the bias and efficiency problems that arise in finite samples.

[6]Recent studies, see for instance Brandt and Kang (2004) and Pástor and Stambaugh (2006), use filtering

We then decompose historical stock returns into shocks to expected returns, the two components of expected dividend growth rates, and unexpected dividend growth rates. The current approach in the literature is to use variance decompositions that have been introduced by Campbell (1991).[7] Our approach has two main advantages over the standard variance decomposition technique. First, our approach does not rely on prespecified instruments. These instruments are required in the VAR approach to form expectations for future returns, which makes this approach sensitive to an omitted variables and a selection bias, and may generate spurious results, see for instance Chen and Zhao (2006). Second, our present-value model is exact, and does not rely on the Campbell and Shiller (1988) approximation.[8] This is particularly important in filtering expected returns and expected growth rates.[9]

Analyzing return decompositions observation-by-observation reveals the underlying economic drivers of the price formation process. For example, we find that the stock market decline at the beginning of the millennium was caused by a negative unexpected dividend growth shock in 2001, followed by a sharp increase in expected returns in 2002. Similarly, the run-up in prices in the second half of the 1990s was due to subsequent positive cash flow surprises and a gradual decline in expected returns. Finally, the oil crises in 1973 and 1974 was accompanied by negative cash flow surprises and an increase in expected returns, leading to sharply negative returns. Furthermore, we find that expected return shocks are generally offset, at least partially, by shocks to the persistent component of expected dividend growth rates.

We can formally test for the presence of a highly persistent predictable component in expected dividend growth, as has been suggested in the recent work by Bansal and Yaron (2004). Interestingly, we cannot formally reject the presence of such a persistent component (the p-value equals 0.087). As a counterfactual, we also consider the case in which the persistent component of dividend growth rates is at the upper boundary of the 95% confidence region. If we estimate our model under this condition, which cannot be rejected statistically, we find that unusual rise in the price-dividend ratio in the 1990s is not only attributable to a decrease in expected returns, but also due to an increase in expected growth rates. An investor who would have acted subject to these beliefs would have outperformed an investor who instead relied on standard predictive regressions.

In the first part of the paper, we develop a tractable closed-form present-value model that

techniques to recover expected returns, but not within a present-value model.

[7] See for instance Campbell and Ammer (1993), Campbell and Vuolteenaho (2004), Cochrane (2006), and Larrain and Yogo (2007).

[8] See, for example, Rytchkov (2007) who employs the Campbell and Shiller (1988) approximation to estimate expected returns and expected growth rates.

[9] Fernández-Villaverde, Rubio-Ramírez, and Santos (2006) show that second-order errors in the transition or measurement equation translate into first-order errors in the likelihood that is used for estimation and filtering.

features time variation in both expected returns and expected dividend growth rates. Our present-value model relates to the linearity-inducing class recently generalized by Gabaix (2007).[10,11] Both expected returns and expected growth rates can follow autoregressive processes of any order. By adding a small twist away from linear autoregressive processes, we obtain a present-value model in which the price-dividend ratio is an exact linear function of expected returns and expected growth rates. We derive the exact likelihood of our model, which allows researchers and practitioners to estimate this class of models by means of maximum likelihood.[12] This procedure avoids several pitfalls that have been highlighted in the predictability literature.[13] First, we can recover the time series of expected returns and expected dividend growth rates, without requiring instruments that are likely to be misspecified.[14] Second, our model delivers an exact relationship between the price-dividend ratio and both expected returns and expected dividend growth. This enables us to avoid log-linear approximations (Campbell and Shiller (1988)), which potentially induce large errors in the filtered series (Fernández-Villaverde, Rubio-Ramírez, and Santos (2006)). In fact, we can exactly compute the higher-order terms that are omitted in the linearized version of our model, and show that these effects can be economically large. Third, we show that imposing the relations implied by the present-value model mitigates the bias and efficiency problems that have received substantial attention in the return predictability literature (Stambaugh (1999) and Lewellen (2004)). Fourth, our present-value approach allows us to effectively handle the errors-in-variables problem that plagues the standard predictive regression approach (Fama and French (1988)). We show that the errors-in-variables problem is most pronounced for predicting future dividend growth rates. Correcting for this problem uncovers a large predictable component of dividend growth rates. Fifth, likelihood-based estimation imposes discipline in selecting and weighing the moment conditions that need to be matched by the model.

[10]Menzly, Santos, and Veronesi (2004) also develop a closed-form present-value model that resides within the linearity-inducing class in a consumption-based general equilibrium model.

[11]An important advantage of our present-value model is that it is linear in expected returns and the expected growth rate of dividends, and it allows us to accommodate stochastic volatility for the innovations as well as flexible error distributions. This freedom in modeling second moments and the distribution of innovations stands in contrast with the present-value models proposed in Burnside (1998) and Ang and Liu (2004), which rely on normality. Ang and Liu (2006) do allow for stochastic volatility in the innovations in a continuous-time present-value model. We, instead, develop a discrete-time present-value model which potentially features stochastic volatility and flexible error distributions.

[12]Our model is non-linear in nature, but we solve it without approximation and use the tractable unscented Kalman filter to construct the likelihood. We compare this non-linear filter to the (exact) particle filter, and find very similar results (Fernández-Villaverde and Rubio-Ramírez (2004), Fernández-Villaverde and Rubio-Ramírez (2006)). Appendix 5.B provides both the theoretical background, and a practical introduction to the particle filter and the unscented Kalman filter.

[13]See Nelson and Kim (1993), Stambaugh (1999), Ang and Bekaert (2007), Ferson, Sarkissian, and Simin (2003), Valkanov (2003), Amihud and Hurvich (2004), Campbell and Yogo (2006), Lewellen (2004), Torous, Valkanov, and Yan (2004), Eliasz (2005), Viceira (1996), Lettau and van Nieuwerburgh (2006), Wachter and Warusawitharana (2007), and Pástor and Stambaugh (2006).

[14]Ferson, Sarkissian, and Simin (2003), Pástor and Stambaugh (2006), and Chen and Zhao (2006).

In addition, the use of Bayesian inference in asset pricing has been popularized recently as it allows one to incorporate prior views on the model specification or model parameters.[15] We show how to compute the likelihood of our model, which makes it straightforward to compute the posterior of the parameters and adopt a Bayesian perspective instead.[16]

Our paper is most closely related to the recent literature on present-value models, see Lettau and van Nieuwerburgh (2006), Pástor and Veronesi (2003), Pástor and Veronesi (2006), Bekaert, Engstrom, and Grenadier (2001), Bekaert, Engstrom, and Grenadier (2005), Ang and Liu (2004), and Brennan and Xia (2005). All of these papers provide closed-form or approximate closed-form expressions for the price-dividend ratio, or the market-to-book ratio in case of firm-level models. However, in case of Bekaert, Engstrom, and Grenadier (2001), Pástor and Veronesi (2003), Pástor and Veronesi (2006), Ang and Liu (2004), and Brennan and Xia (2005) the price-dividend ratio is an infinite sum, or indefinite integral, of exponentially quadratic terms, which makes likelihood-based estimation and filtering computationally intractable. Bekaert, Engstrom, and Grenadier (2001) and Ang and Liu (2004) estimate the model by means of GMM and model expected returns and expected growth rates as an affine function of a set of prespecified instruments. Also, the model requires innovations to expected returns and expected growth rates to be (conditionally) Gaussian. Brennan and Xia (2005) use a two-step procedure to estimate their model and use long-term forecasts for expected returns to recover an estimate of the time series of (instantaneous) expected returns, in turn. Alternatively, Lettau and van Nieuwerburgh (2006) set up a linearized present-value model and recover structural parameters from reduced-form estimators. They subsequently test whether the present-value constraints are violated. Lettau and van Nieuwerburgh (2006) impose, however, that the persistence of expected returns and expected growth rates is equal. Our main theoretical contributions are the following. First, we develop a tractable closed-form present-value model that allows for (i) variation in expected returns and expected dividend growth rates with multiple frequencies and (ii) non-Gaussian error distributions. Second, we provide a framework for likelihood-based estimation and filtering without relying on prespecified instruments. Using this economic model, we find that the price-dividend ratio strongly predicts both future returns and future dividend growth rates.

The paper proceeds as follows. In Section 2, we develop our closed-form present-value model. In addition, we discuss the econometric framework to estimate present-value models. Section 3 contains our main empirical results on why prices and price-dividend ratios move

[15] Important recent contributions include Cremers (2002), Avramov (2002), Avramov (2004), Avramov and Chordia (2006), Pástor (2000), and Wachter and Warusawitharana (2007).

[16] Our present-value model features considerable non-linearities, which implies that the Gibbs sampler cannot be used easily. However, Fernández-Villaverde and Rubio-Ramírez (2006) show how to combine the Metropolis-Hastings algorithm with the particle filter, which is one of the two non-linear filtering procedures we implement.

over time. Section 4 studies the predictability of dividend growth rates in more detail. In particular, we allow for more flexible dynamics of expected growth rates and formally test for the presence of a highly persistent expected dividend growth component. Section 5 contains several extensions of the standard models. Section 6.10 concludes.

5.2 Present-value model

5.2.1 Theoretical model

We develop in this section a present-value model that results in a closed-form expression for the price-dividend ratio. Let P_t denote the cum-dividend price of the stock at time t, and D_t the time-t dividend. We adopt this pricing convention purely for analytical convenience. The price-dividend ratio then reads:

$$PD_t \equiv \frac{P_t}{D_t}, \tag{5.2}$$

and simple stock returns (gross) are given by:

$$R_{t+1} = \frac{P_{t+1}}{P_t - D_t} = \frac{PD_{t+1}}{PD_t - 1}\frac{D_{t+1}}{D_t}. \tag{5.3}$$

We define the discount rate, or expected return, at time t as $\mu_t^\star$:

$$\mu_t^\star \equiv E_t\left[R_{t+1}\right] - 1. \tag{5.4}$$

Combining (5.3) and (5.4) leads to the well-known present-value relation:

$$P_t = D_t + \frac{E_t\left[P_{t+1}\right]}{1+\mu_t^\star}. \tag{5.5}$$

To complete the specification of the present-value model, we need to specify the dynamics of dividend growth rates, D_{t+1}/D_t, and expected returns, $\mu_t^\star$. We model dividend growth as:

$$\begin{aligned} \frac{D_{t+1}}{D_t} &= (1+g_t)(1+\varepsilon_{t+1}^D) \\ &= 1 + g_t + (1+g_t)\varepsilon_{t+1}^D, \end{aligned} \tag{5.6}$$

where g_t indicates the expected dividend growth rate. The multiplicative form in (5.6) is the first element of the linearity-inducing model.[17] The small deviation from the standard

[17] Gabaix (2007) presents the linearity-inducing class for pricing kernels, which requires one to make an assumption on which shocks are priced, and which prices of risk are time varying in turn. Our goal is

(homoscedastic) autoregressive model is, however, for all practical purposes inconsequential. The second step is to introduce the variable μ_t that is approximately equal to the discount factor $\mu_t^\star$. We define μ_t as:

$$\frac{1+g_t}{1+\mu_t^\star} \equiv 1 + g_t - \mu_t. \tag{5.7}$$

Up to a first-order approximation, $\mu_t^\star$ and μ_t will be the same.[18] However, since we do not want to rely on approximations, we introduce μ_t whose dynamics we model in turn. We refer to μ_t and $\mu_t^\star$ as expected returns interchangeably, having noted the difference in equation (5.7). We decompose expected growth rates and expected returns as:

$$g_t = \gamma_0 + \hat{g}_t, \tag{5.8}$$
$$\mu_t = \delta_0 + \hat{\mu}_t, \tag{5.9}$$

where $\hat{g}_t$ and $\hat{\mu}_t$ follow a (near-)AR(1) process:

$$\hat{g}_{t+1} = \gamma_{1t}\hat{g}_t + \varepsilon_{t+1}^g, \tag{5.10}$$
$$\hat{\mu}_{t+1} = \delta_{1t}\hat{\mu}_t + \varepsilon_{t+1}^\mu. \tag{5.11}$$

Hence, discount rates and expected growth rates follow de-meaned AR(1) processes, apart from the linearity-inducing twist in γ_{1t} and δ_{1t}. The decomposition in (5.8) and (5.9) is purely for analytical convenience and the constants δ_0 and γ_0 will be estimated jointly with all other parameters. In the standard VAR approach, γ_{1t} and δ_{1t} would be constant parameters. We consider instead:

$$\gamma_{1t} = \lambda_t \gamma_1, \tag{5.12}$$
$$\delta_{1t} = \lambda_t \delta_1, \tag{5.13}$$

in which λ_t and α are given by:

$$\lambda_t \equiv \frac{\alpha}{\alpha + \hat{g}_t - \hat{\mu}_t}, \tag{5.14}$$
$$\alpha \equiv 1 + \gamma_0 - \delta_0. \tag{5.15}$$

In other words, both discount rates and growth rates follow an AR(1) process, apart from the fact that the autoregressive parameters, γ_{1t} and δ_{1t}, are slightly time varying. The degree

different. We, instead, model the expected return directly, without making explicit assumptions on the prices risk of each of the shocks.

[18] Empirically, we find that the correlation between the two series is 0.998.

of time variation is captured by a common factor, λ_t, defined in (5.14). The process λ_t will hover closely around one and the unconditional means of g_t and μ_t will be very close to γ_0 and δ_0, respectively. To summarize, by introducing μ_t in (5.7) and λ_t in (5.14) we twist the standard affine model. This results in a tractable and closed-form expression for the price-dividend ratio as we show below.

We consider the following structure for the innovations. The covariance between innovations in unexpected dividend growth and expected returns is given by: $\text{Cov}(\varepsilon^D_{t+1}, \varepsilon^\mu_{t+1}) = \sigma_{D\mu}$. Likewise, the covariance between innovations in expected growth rates and expected returns is given by: $\text{Cov}(\varepsilon^g_{t+1}, \varepsilon^\mu_{t+1}) = \sigma_{g\mu}$. To ensure statistical identification of the model, we assume that innovations in expected growth rates and unexpected growth rates are uncorrelated, $\text{Cov}(\varepsilon^g_{t+1}, \varepsilon^D_{t+1}) = \sigma_{gD} = 0$.[19]

Appendix A shows that, under the assumptions stated above, the price dividend-ratio is an affine function of the de-meaned expected dividend growth rate ($\hat{g}_t$) and the de-meaned expected return or discount rate ($\hat{\mu}_t$):

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t, \tag{5.16}$$

where A, B_1, and B_2 are constants given by:

$$A = \left(1 - \alpha + \frac{\sigma_{D\mu}\alpha}{1 - \delta_1\alpha + \sigma_{D\mu}}\right)^{-1}, \tag{5.17}$$

$$B_1 = \frac{-A}{1 - \delta_1\alpha + \sigma_{D\mu}}, \tag{5.18}$$

$$B_2 = \frac{A + B_1\sigma_{D\mu}}{1 - \gamma_1\alpha}, \tag{5.19}$$

and recall that α is given by $\alpha = 1 + \gamma_0 - \delta_0$.

We develop the main intuition for the model by considering three special cases. First, suppose that expected returns and expected growth rates are constant. This is achieved by setting $\delta_1 = \gamma_1 = 0$, as well as $\sigma_\mu = \sigma_g = 0$. In that case, the price-dividend ratio simplifies to:

$$PD = A = (1 - \alpha)^{-1} = \frac{1}{\delta_0 - \gamma_0}, \tag{5.20}$$

which is the standard Gordon-growth formula.

[19] A similar assumption is often made in modeling the dynamics of inflation. In that case, expected inflation is assumed to follow an AR(1)-process. The resulting process for inflation is an ARMA(1,1)-process. The correlation between expected and unexpected inflation cannot be identified (see Campbell and Viceira (2001a) and Sangvinatsos and Wachter (2005)), and is often normalized to zero.

As a second special case, suppose that shocks to expected returns and shocks to unexpected dividend growth are uncorrelated, i.e., $\sigma_{D\mu} = 0$. In this case, the coefficients simplify to:

$$A = \frac{1}{\delta_0 - \gamma_0}, \tag{5.21}$$

$$B_1 = \frac{-A}{1 - \delta_1 \alpha}, \tag{5.22}$$

$$B_2 = \frac{A}{1 - \gamma_1 \alpha}. \tag{5.23}$$

The coefficient A still represents the unconditional average and is the same as in (5.20). However, the price-dividend ratio will now move in response to changes in expected returns or expected dividend growth rates. The impact on the price-dividend ratio is captured by B_1 and B_2, respectively. For plausible parameter values, A will be positive and α near 1. This implies that B_1 is negative, and B_2 is positive. Considering the return definition in (5.3), this implies that an increase in expected returns leads, ceteris paribus, to a negative contemporaneous return, while a positive shock to growth rates induces a positive contemporaneous returns. The persistence parameters of discount rates and expected growth rates (δ_1 and γ_1) determine their quantitative impact on the price-dividend ratio.

We finally consider the special case where expected growth rates are constant. This special case has often been considered in the extant literature. In this case, the price-dividend ratio then reduces to:

$$PD_t = A + B_1 \hat{\mu}_t, \tag{5.24}$$

in which A and B_1 are given in (5.17) and (5.18). The otherwise unobservable expected return, μ_t, can now be written as an affine function of the observable price-dividend ratio:

$$\mu_t = \delta_0 + \frac{PD_t - A}{B_1}. \tag{5.25}$$

This implies that, apart from the fact that the coefficients need to be estimated, the observed price-dividend ratio can be used directly to recover the time series of expected returns.

5.2.2 Why do prices move?

Our model provides an exact (i.e., not approximated) answer to the question why historical prices, returns, and price-dividend ratios moved. In a present-value model, any change in these variables is attributable to changes in expected returns, expected growth rates, or unexpected growth rates. These changes may be expected or unexpected. The expected

change of the price-dividend ratio is simply given by:

$$E_t\left[PD_{t+1}\right] - PD_t \;=\; B_1\left(\delta_{1t}-1\right)\hat{\mu}_t + B_2\left(\gamma_{1t}-1\right)\hat{g}_t, \tag{5.26}$$

and the unexpected change of the price-dividend ratio is given by:

$$PD_{t+1} - E_t\left[PD_{t+1}\right] \;=\; B_1\varepsilon^{\mu}_{t+1} + B_2\varepsilon^{g}_{t+1}. \tag{5.27}$$

Similarly, we can decompose the realized stock return into the conditional expected stock return, $\mu^{\star}_t \equiv E_t\left(R_{t+1}\right) - 1$, and the unexpected return as:

$$R_{t+1} - \left(1+\mu^{\star}_t\right) = h_t \left[\begin{array}{c} B_1\varepsilon^{\mu}_{t+1} + B_2\varepsilon^{g}_{t+1} + E_t\left[PD_{t+1}\right]\varepsilon^{D}_{t+1} \\ +B_1\left(\varepsilon^{D}_{t+1}\varepsilon^{\mu}_{t+1} - \sigma_{D\mu}\right) + B_2\varepsilon^{D}_{t+1}\varepsilon^{g}_{t+1} \end{array} \right], \tag{5.28}$$

where:

$$h_t \equiv \frac{\left(1+g_t\right)}{PD_t - 1}. \tag{5.29}$$

This return decomposition has a straightforward interpretation. A positive shock to the discount factor of one percentage point will, ceteris paribus, lead to a negative realized return of $0.01 \times B_1 h_t$. Note that for reasonable parameter values, the coefficient B_1 is negative. Similarly a positive shock to the expected dividend growth rate of one percentage point will lead to a positive realized return equal to $0.01 \times B_2 h_t$. Finally, a positive unexpected dividend shock of one percentage point will lead to a positive realized return of $0.01 \times E_t\left(PD_{t+1}\right) h_t$. The last two terms (i.e., the multiplications of shocks) in equation (5.28) are an order of magnitude smaller than the first three terms. This return decomposition provides an alternative to Campbell (1991) within our closed-form present-value model. Note that the return decomposition of Campbell (1991) is based on geometric returns and relies on a first-order approximation, while we perform an exact decomposition of arithmetic returns instead.

If we assume in addition that $(\varepsilon^{g}_{t+1}, \varepsilon^{\mu}_{t+1}, \varepsilon^{D}_{t+1})$ are jointly normal,[20] it is straightforward to show that the conditional variance of stock returns is given by:

$$\begin{array}{rcl} \sigma^2_{R,t} & \equiv & E_t\left[R_{t+1} - E_t\left[R_{t+1}\right]\right]^2 \\ & = & h_t^2 \left(\begin{array}{c} B_1^2\sigma^2_{\mu} + B_2^2\sigma^2_{g} + \left(E_t\left[PD_{t+1}\right]\right)^2\sigma^2_{D} + B_1^2\left(\sigma^2_{\mu}\sigma^2_{D} + 2\sigma^2_{D\mu}\right) \\ +B_2^2\sigma^2_{g}\sigma^2_{D} + B_1B_2\sigma_{g\mu} + B_1E_t\left[PD_{t+1}\right]\sigma_{D\mu} + B_1B_2\sigma^2_{D}\sigma_{g\mu} \end{array} \right), \end{array} \tag{5.30}$$

[20] Note that up until this point we have not made any distributional assumptions on the error terms, apart from their covariance structure and their zero means.

in which we use results in Haldane (1942) to compute the moments of the product of powers of normally distributed innovations. Equation (5.30) shows that even when the underlying shocks of the model are homoscedastic, returns will be heteroscedastic. The present-value model therefore endogenously generates stochastic volatility in stock returns (Ang and Liu (2006)).[21]

The conditional covariance between expected returns and unexpected returns is given by:

$$\mathrm{Cov}_t\left(R_{t+1} - E_t\left[R_{t+1}\right], \varepsilon_{t+1}^{\mu}\right) = h_t\left(B_1\sigma_\mu^2 + B_2\sigma_{\mu g} + E_t\left[PD_{t+1}\right]\sigma_{\mu D}\right) \tag{5.31}$$

The conditional correlation between expected returns and unexpected returns can then easily be computed by dividing the covariance above by σ_μ and $\sigma_{R,t}$. This correlation has received a lot of attention in the literature. It is exactly this correlation which generates the bias in standard linear predictive regressions, see Stambaugh (1999) for instance. We show that this correlation is a function of the underlying parameters of the present-value model. A similar result has been derived in Cochrane (2006) and Lettau and van Nieuwerburgh (2006) for geometric returns and under the Campbell and Shiller (1988) approximation.

Pástor and Stambaugh (2006) assume a prior distribution over this correlation and they note that it is likely to be negative. In our model, this correlation can be computed exactly, and we show how it depends on the structural parameters of the present-value model. When both $\sigma_{g\mu}$ and $\sigma_{D\mu}$ are zero, this correlation is negative. Its magnitude then depends on the variance decomposition of returns. If expected growth rates are constant and unexpected dividend growth is not too volatile, then realized returns are mainly driven by discount rate shocks and the correlation will be close to minus one. However, if returns are mainly driven by cash-flow shocks (either expected or unexpected), then the correlation will be closer to zero. Furthermore, in case $\sigma_{g\mu}$ and/or $\sigma_{D\mu}$ are positive, theoretically speaking the sign of this correlation can be both positive and negative. This is not surprising. If positive shocks to the discount factor go hand in hand with positive shocks to dividend growth (expected or unexpected) the net effect on returns is ambiguous. An important determinant then is the relative persistence of the shocks.

Finally, note that it is also the correlation between expected and unexpected returns that drives the hedging demands and gains to dynamic investment in long-run portfolio choice problems. If the correlation is strongly negative, then bad return realizations are ameliorated by better future return prospects, which motivates the investor to put a higher weight in stocks. However, if the correlation is close to zero, stocks do not form a good hedge against future investment opportunities, which reduces the overall allocation to stocks. Our model

[21] With mild assumptions on the correlation structure, our model can easily accommodate stochastic volatility for all three shocks.

implies that this correlation can be time varying, implying that hedging demands will be more important in some periods than others.[22]

5.2.3 Alternative present-value models

As an alternative to our present-value model, Burnside (1998), Pástor and Veronesi (2003), Pástor and Veronesi (2006), Bekaert, Engstrom, and Grenadier (2001), Bekaert, Engstrom, and Grenadier (2005), Ang and Liu (2004), and Brennan and Xia (2005) derive a closed-form expression for the price-dividend ratio or market-to-book ratio as well. This closed-form expression is an infinite sum, or indefinite integral in case of continuous-time models, of exponentially quadratic functions. This makes non-linear filtering virtually impossible from a computational perspective. Our model also delivers a closed-form expression for the price-dividend ratio that is, in contrast, affine in expected returns and growth rates. This makes the filtering problem computationally highly tractable.

5.3 Data and econometric approach

In this section, we estimate the present-value model developed in Section 5.2. First, Section 5.3.1 provides details on the data used in estimation. In Section 5.3.2, we consider the case in which expected growth rates are constant so that we need no filtering techniques to recover the time series of expected returns. We use this model to study the properties of our estimator in detail, and compare it to the more conventional predictive regressions approach. In Section 5.3.2, we discuss the estimation of the full model and outline the filtering procedure used to recover the time series of expected returns and growth rates.

5.3.1 Data

We create a series of effectively paid out dividends, D_t, and use the cum-dividend price levels to construct the cum-dividend price-dividend ratio, PD_t. The cum-dividend price is the ex-dividend price at the end of year t augmented with the effective dividends D_t paid out during year t. The data is constructed using the total payout yield by Boudoukh, Michaely, Richardson, and Roberts (2004). There is an increasing amount of evidence that firms prefer stock repurchases over dividend payments (Brav, Graham, Harvey, and Michaely (2005)) and that dividends are concentrated in fewer firms (Skinner (2006)). We therefore consider the

[22]Buraschi, Porchia, and Trojani (2007) study the impact of time-varying correlations and volatilities on optimal portfolio choice. They show that the hedging demand that is induced by time-varying correlations constitutes a non-negligible part of the optimal portfolio, and often dominates the hedging demands caused by time-varying volatilities.

total payout data to be the relevant source of data.[23] We estimate the model with annual data over the period 1946-2005, i.e., we focus on the post-war sample.

5.3.2 Likelihood-based estimation

Constant expected growth rates

To illustrate how our model can help improve the finite sample properties of the predictive coefficients, we first consider the special case in which expected growth rates are constant, i.e., $\gamma_1 = \sigma_g = 0$. At least, in this case a standard predictive regression does not suffer from an errors-in-variables problem. However, we show that a standard predictive regression is highly inefficient and induces a small-sample bias that is much larger compared to estimating the present-value model directly by maximum likelihood. Recall that in the case where expected dividend growth rates are constant, the price-dividend ratio is given by:

$$PD_t = A + B_1 \hat{\mu}_t. \tag{5.32}$$

We observe data on returns, $R^T \equiv \{R_1, \ldots, R_T\}$, dividend growth rates, $\Delta D^T \equiv \{D_1/D_0, \ldots, D_T/D_{T-1}$ and price-dividend ratios, $PD^T \equiv \{PD_0, \ldots, PD_T\}$, which we in turn use to estimate the structural model parameters, i.e., $\Theta \equiv \{\delta_0, \delta_1, \gamma_0, \sigma_D, \sigma_\mu, \sigma_{D\mu}\}$. The key observation is that the time series of returns follows from the time series of dividend growth and the price-dividend ratio (see also Cochrane (2006) and Lettau and van Nieuwerburgh (2006)). As such, separately including the returns series in estimation does not add any information. We thus estimate our model using dividend growth and price-dividend ratio data only.

We employ maximum-likelihood estimation, which accounts for all relationships implied by the present-value model. We define the innovations $\xi_{t+1} \equiv (B_1 \varepsilon^{\mu}_{t+1}, (1+\gamma_0)\varepsilon^{D}_{t+1})'$:

$$\begin{aligned} \xi_{1,t+1}(\gamma_0, \delta_0, \delta_1) &= PD_{t+1} - A - B_1 \delta_{1t} \hat{\mu}_t, & (5.33) \\ \xi_{2,t+1}(\gamma_0) &= \frac{D_{t+1}}{D_t} - 1 - \gamma_0, & (5.34) \end{aligned}$$

where δ_{1t} depends on α, and α, A, and B_1 are functions of γ_0, δ_0, $\sigma_{D\mu}$, and δ_1, as given by (5.15), (5.17) and (5.18). Note that when deriving the results above, we have not yet made any assumptions about the distribution of the error term. For estimation purposes,

[23] We also estimated our model using CRSP data without explicitly accounting for share repurchases. Our qualitative results are not affected. The same comment applies to the sample period from 1927 to 2005 and when we use real prices and dividends instead of their nominal counterparts.

we assume:

$$\xi_{t+1} \sim N\left(0_{2\times 1}, \begin{pmatrix} B_1^2\sigma_\mu^2 & (1+\gamma_0)B_1\sigma_{D\mu} \\ (1+\gamma_0)B_1\sigma_{D\mu} & (1+\gamma_0)^2\sigma_D^2 \end{pmatrix}\right) = N(0_{2\times 1}, \Sigma). \tag{5.35}$$

The (scaled) conditional log-likelihood then reads:

$$\mathcal{L}\left(\Theta; \Delta D^T, PD^T\right) = -T \log \mid \Sigma \mid - \sum_{t=1}^{T} \xi_t' \Sigma^{-1} \xi_t. \tag{5.36}$$

At first sight it may seem counter-intuitive that we are estimating the predictable variation in expected returns without using actual returns in our likelihood. However, as mentioned above, it is important to note that observing dividend growth and the price-dividend ratio is equivalent to observing returns.

Comparison with predictive regressions

We propose to estimate our model with maximum likelihood. One possible way to estimate (part of) this model is to run a standard predictive regression, which implies regressing realized returns on the lagged price-dividend ratio:

$$R_{t+1} = \alpha + \beta PD_t + \eta_{t+1}, \tag{5.37}$$

see for instance Cochrane (2006), Campbell and Yogo (2006), Campbell and Thompson (2007), Goyal and Welch (2003), Goyal and Welch (2006), and Lettau and van Nieuwerburgh (2006). Ignoring the difference between μ_t and $\mu_t^\star$, we can link (5.37) to our structural present-value model (5.32):

$$0 = \alpha + \beta A, \tag{5.38}$$

$$1 = \beta B_1. \tag{5.39}$$

The parameter β has been studied widely in the return predictability literature. Due to the empirically high negative correlation between innovation in the price-dividend ratio and the high persistence of the price-dividend ratio, β tends to be downward biased (Stambaugh (1999) and Amihud and Hurvich (2004)). To focus on the small-sample properties of the different estimators, we present a comparison of estimation techniques for the case where expected dividend growth rates are constant. We then show how maximum-likelihood estimation reduces the small-sample bias and improves efficiency. We do not compare estimators for the case where expected growth rates are time varying as well, because in this case the price-dividend ratios can move due to changes in expected returns and changes in

expected growth rates. This leads to an errors-in-variables bias, which affects the parameter β not only in small samples, but also in population (Fama and French (1988), Kothari and Shanken (1992), Goetzman and Jorion (1995), and Koijen and Nieuwerburgh (2007)). This errors-in-variables problem is particularly important when expected dividend growth rates are persistent, as we find in Section 5.5.

To this end, we consider the following set of parameters that are close to the maximum-likelihood estimates:

$$\delta_0 = 0.06,\ \delta_1 = 0.85,\ \gamma_0 = 0.025,\ \sigma_D = 0.14,\ \sigma_\mu = 0.018,\ \sigma_{D\mu} = 0.00013. \tag{5.40}$$

We simulate 10,000 series of 60 annual observations. For each of these series, we run the predictive regression described in (5.37) and, as an alternative, we estimate the model by means of maximum likelihood. We are interested in the distribution of the predictive coefficient, β, that follows from both estimation procedures. One might argue that we are favoring the maximum-likelihood estimation as we approximate $\mu_t \simeq \mu_t^\star$. To address this concern, we also perform the following predictive regression:

$$R_{t+1} = \hat{\alpha} + \hat{\beta} x_t + \hat{\varepsilon}_{t+1}, \tag{5.41}$$

where $x_t = B_1 \mu_t^\star = B_1 \left[(1+\gamma_0)/(\alpha - B_1^{-1}(PD_t - A)) - 1\right]$. For this transformation of the price-dividend ratio, we have $\hat{\beta} = B_1^{-1}$. We will call this latter estimator the adapted OLS estimator. The results are presented in Table 5.1 and Figure 5.1. The top panel of Figure 5.1 depicts the distribution of the OLS estimator, the middle panel of the adapted OLS estimator, and the bottom panel of our maximum-likelihood estimator.

The main results can be summarized as follows. First, the maximum-likelihood estimator, which takes into account the full present-value model, is much more efficient and less biased. In this particular experiment, the root mean squared error (RMSE) of the maximum-likelihood estimator is 45 per cent lower than the RMSE of the OLS estimator and 35 per cent lower than the RMSE of the adapted OLS estimator. Such efficiency gains are equivalent to having three to four times as many observations. Part of this efficiency gain is due the fact that all present-value models imply a particular form of heteroscedasticity (see Ang and Liu (2006)). Taking this heteroscedasticity into account can improve efficiency. There is, however, another important source of efficiency gain. The correlation between expected and unexpected returns is a function of the underlying present value parameters as explained in equation (5.31). Imposing this functional relationship in estimation further improves efficiency. Second, comparing the OLS and the adapted OLS estimator in Table 5.1 illustrates the importance of fully taking into account the non-linearities of the present-value model.

The adapted OLS estimator, which corrects for the non-linearities in the model, performs much better than the regular OLS estimator, which can be interpreted as a linearization of the adapted OLS estimator. This indicates that non-linearities play an important role in present-value models and that first-order approximations can substantially deteriorate the estimates and filtered series (Fernández-Villaverde, Rubio-Ramírez, and Santos (2006)).

Time-varying expected growth rates

We now consider the general case in which both discount rates and expected growth rates are allowed to be time varying. In this case, a standard predictive regression of either returns or dividend growth rates on the lagged price-dividend ratio suffers from an error-in-variables problem, which does not disappear as the sample size increases. Our maximum likelihood approach resolves this error-in-variables problem.

Recall that in Section 5.3.2, expected returns are an affine function of the price-dividend ratio, which, up to estimation error of the structural parameters of the present-value model, makes it straightforward to uncover the former. However, now the observable series of price-dividend ratios, PD_t, is an affine function of two latent processes (see (5.16)), namely $\hat{\mu}^T$ and $\hat{g}^T$. Hence, we observe two time series (PD^T and ΔD^T), and have three series of innovations in the model ($\varepsilon^{g,T}$, $\varepsilon^{D,T}$, and $\varepsilon^{\mu,T}$). It is therefore no longer possible to exactly recover the latent time series, and we need to apply filtering techniques to estimate the time series of expected returns and growth rates instead.

We implement two non-linear filters to perform filtering and estimation in our model, namely the particle filter and the unscented Kalman filter. We refer to Doucet, de Freitas, and Gordon (2001) for a text book treatment of the particle filter. Fernández-Villaverde and Rubio-Ramírez (2004) and Fernández-Villaverde and Rubio-Ramírez (2006) use the particle filter to estimate dynamic stochastic general equilibrium (DSGE) models in macro economics. Even though the particle filter is straightforward to implement, it is computationally more costly than, for instance, the extended Kalman filter. This raises the question why we cannot simply linearize our model, for instance using the Campbell and Shiller (1988) approximation, and use the extended Kalman filter instead (Jazwinski (1973)). Fernández-Villaverde, Rubio-Ramírez, and Santos (2006) show however that second-order errors in the measurement and transition equations lead to first-order errors in the likelihood that we want to use in turn for estimation and filtering. Fernández-Villaverde and Rubio-Ramírez (2004) and Fernández-Villaverde and Rubio-Ramírez (2006) show empirically that the resulting filtered series can look dramatically different. Our closed-form present-value model allows us to solve the filtering problem without any approximation, which is required for likelihood-based inference. As an alternative, we also employ the unscented Kalman filter. The unscented Kalman filter

is second-order accurate (see Fontaine and Garcia (2007), Bakshi, Carr, and Wu (2007), and Wan and van der Merwe (2001)) in non-linear models. We find similar results for the unscented Kalman filter and the particle filter, but the former is computationally much faster. It takes approximately 10 seconds to estimate our model on a standard desktop computer using Matlab. Appendix 5.B provides a detailed description of the particle filter and the unscented Kalman filter, as well as a practical guide to implementation.

We provide here further details on the transition and measurement equations that characterize the filtering problem. The transition equations in our model are given by:

$$\hat{g}_{t+1} = \gamma_{1t}\hat{g}_t + \varepsilon^g_{t+1}, \tag{5.42}$$
$$\hat{\mu}_{t+1} = \delta_{1t}\hat{\mu}_t + \varepsilon^\mu_{t+1}, \tag{5.43}$$

and the measurement equations using data PD^T and ΔD^T:

$$\frac{P_{t+1}}{D_{t+1}} = A + B_1\hat{\mu}_{t+1} + B_2\hat{g}_{t+1}, \tag{5.44}$$
$$\frac{D_{t+1}}{D_t} = 1 + \hat{g}_t + (1+\hat{g}_t)\varepsilon^D_{t+1}. \tag{5.45}$$

We can, however, reduce the number of latent variables to one using the relation of the price-dividend ratio to expected returns and growth rates in (5.44). We then have only one transition equation:

$$\hat{g}_{t+1} = \gamma_{1t}\hat{g}_t + \varepsilon^g_{t+1}, \tag{5.46}$$

and two measurement equations:

$$\frac{D_{t+1}}{D_t} = 1 + \hat{g}_t + (1+\hat{g}_t)\varepsilon^D_{t+1}, \tag{5.47}$$
$$PD_{t+1} = A + \delta_{1t}\left[PD_t - A - B_2\hat{g}_t\right] + B_2\gamma_{1t}\hat{g}_t + B_1\varepsilon^\mu_{t+1} + B_2\varepsilon^g_{t+1} \tag{5.48}$$

In the filtering approach, we need to make further assumptions about the distribution of the innovations.[24] We assume all innovations to be normally distributed with a constant covariance matrix. The correlation structure between the innovations is as discussed in Section 5.2, and accommodates non-zero correlation between expected returns and expected growth rates, and between expected returns and unexpected growth rates. We do note, though, that these distributional assumptions and constant second moments are not a requirement

[24]It is straightforward to provide formal conditions to ensure the stability of linearity-inducing processes, see Gabaix (2007). Intuitively, the process for expected returns needs to be bounded from above, and the process for expected growth rates needs to bounded from below. This also ensures that the price-dividend ratio remains positive. We find, however, that the regularity conditions are easily satisfied in our empirical work.

of the model and that we can easily handle alternative distributions as well as time-varying volatility.

5.4 Empirical results

This section contains the empirical results when expected returns and expected growth rates follow a (near) AR(1) process. Section 5.4.1 presents the estimation results of the present-value model in Section 5.2. Section 5.4.2 decomposes historical returns into fundamental shocks in our present-value model. Likewise, Section 5.4.3 decomposes the historical price-dividend ratio into expected returns and expected growth rates. In Section 5.4.4, we analyze the information content present in the price-dividend ratio to predict future returns and dividend growth rates. Motivated by the results of this section, where expected growth rates exhibit substantial time-variation but little persistence, we allow for an additional frequency in expected growth rates in Section 5.5, which we find to be persistent.

5.4.1 Estimation results

The estimation results are summarized in Table 5.2. Panel A presents the parameter estimates of the processes that govern the dynamics of expected returns, expected growth rates, and realized growth rates, with bootstrapped standard errors between brackets. Panel B then displays the implications for the coefficients of the price-dividend ratio and the constant α.

There are at least four interesting aspects to our estimates that we discuss in turn. First, it turns out that shocks to the discount factor and shocks to expected growth rates are highly positively correlated ($\rho_{g\mu} = 0.79$), see also Lettau and Ludvigson (2005). Second, the autoregressive coefficient of expected growth rates (γ_1) is negative, while the autoregressive coefficient of expected returns (δ_1) is highly positive. Recall that the price-dividend ratio is affine in expected returns and expected growth rates. Our estimates imply that the price-dividend ratio is comprised of a persistent component related to expected returns and a high-frequency component related to expected growth rates. As such, the high-frequency component of the price-dividend ratio predicts future dividend growth, whereas the persistent component predicts future returns. Note also (Panel B) that the coefficient on expected returns in the price-dividend ratio (B_1) is as a result much larger in absolute value than the coefficient on expected growth rates (B_2). Third, unexpected shocks to dividend growth (ε^D_{t+1}) and innovations to expected returns (ε^μ_{t+1}) are virtually uncorrelated. Fourth, most parameters are accurately estimated. For instance, we find that the persistence of expected returns, and the correlation between innovations to expected returns and expected growth rates are both reliably different from zero.

Figure 5.2 portrays our filtered series for expected returns ($\mu_t^\star$) and Figure 5.3 for expected dividend growth rates (g_t). Both figures also display the fitted values following from an OLS regression with the price-dividend ratio as the predictor variable. Finally, we plot the realized values of returns and dividend growth. Figure 5.2 shows that our filtered series of expected returns closely tracks the expected return series following from a standard predictive regression. However, our series is slightly higher in the 1990s. Figure 5.3 shows that there can be large discrepancies between the series following from the present-value model and from a standard predictive regression. The filtered series from the present-value model is much more transient, as γ_1 is negative. The series following from the predictive regression tracks the price-dividend ratio by construction. It is important to note that the coefficient in the predictive regression has the incorrect sign and is insignificant (not reported). From the present-value model, it follows that higher expected growth rates should lead to a higher price-dividend ratio, keeping expected returns constant. This suggests that the errors-in-variables problem is severe in case of the predictive regression for the dividend growth rates.

5.4.2 Why do prices move?

The present-value model allows us to decompose historical stock returns, see (5.28), which we repeat here:

$$R_{t+1} = 1 + \mu_t^\star + h_t \begin{bmatrix} B_1\varepsilon_{t+1}^{\mu} + B_2\varepsilon_{t+1}^{g} + E_t\left[PD_{t+1}\right]\varepsilon_{t+1}^{D} \\ +B_1\left(\varepsilon_{t+1}^{D}\varepsilon_{t+1}^{\mu} - \sigma_{D\mu}\right) + B_2\varepsilon_{t+1}^{D}\varepsilon_{t+1}^{g} \end{bmatrix}. \tag{5.49}$$

Figure 5.4 and Figure 5.5 portray the unexpected return, decomposition for each stock return in our sample period.[25] In both figures, the top panel displays the unexpected returns, and the bottom panels the decomposition. The decomposition consists of three first-order effects, one for each shock as shown in Figure 5.4, and two second-order effects as shown in Figure 5.5. Equation (5.49) shows that these second-order effects make the decomposition exact in our model. Each unexpected return is the result of a shock to discount rates, a shock to expected dividend growth rates, and a shock to unexpected dividend growth rates. As noted before, analyzing the return decomposition observation-by-observation reveals the underlying economic drivers of the price formation process. For instance, Figure 5.4 indicates that the stock market decline at the beginning of the millennium was caused by a negative unexpected dividend growth shock in 2001 (ε_D),[26] followed by a sharp increase of the discount

[25]The unexpected return is the difference between the realized returns and expected returns as portrayed in Figure 5.2.

[26]Our explanation for the downturn of the aggregate stock market resonates with the findings of Pástor and Veronesi (2006). Pástor and Veronesi (2006) argue that the burst in the Nasdaq "bubble" can be largely attributed to an unexpected drop in firm profitability. Their Figure 6 shows that Nasdaq's ROE was -20%

factor in 2002 (ε_μ). Similarly, the run-up in prices in the second half of the 1990s was due to several subsequent positive cash flow surprises and a gradual steady decline of the discount factor. Next, the oil crises in 1973 and 1974 corresponded to negative cash flow surprises and an increase in the discount factor, leading to large negative unexpected returns. Finally, note that the results in Figure 5.5 relative to Figure 5.4 clearly indicate that second-order effects, which are ignored in log-linearizations, are relatively small but most certainly not negligible. This provides an additional motivation to use a closed-form present-value model that does not rely on first-order approximations.

We now consider the variance decomposition of unexpected returns. Equation (5.49) decomposes unexpected stock returns into five components. Table 5.3 provides the covariance matrix of each of these terms. To this end, we use the estimated innovations as in Figure 5.4 and 5.5, i.e., three first-order effects and two second-order effects, and we determine their unconditional covariance. We then scale all elements with the variance of unexpected returns. The resulting fractions in Table 5.3 therefore indicate how each of the terms contributes to the total variance of unexpected stock returns. The sum of all elements in Table 5.3 is 100%.

We find that two main sources of variation in unexpected stock returns are due to changes in expected returns, (1), and unexpected dividend growth, (3). Shocks to expected growth rates have little impact on unexpected stock returns due to low persistence of expected growth rate shocks. Likewise, the second-order terms ((4) and (5)) only marginally contribute to the total unexpected stock return.

Finally, we plot in Figure 5.7 the correlation between expected and unexpected returns, which we derive analytically in equation (5.31). The graph indicates that this correlation is on average negative with a mean of -0.60, but that it substantially varies over time, tracking the variation of the price-dividend ratio. It takes values as high as -0.45 and as low as -0.67.

5.4.3 Why does the price-dividend ratio move?

Recall that the price-dividend ratio in our model is given by:

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t. \tag{5.50}$$

Figure 5.6 decomposes the demeaned price-dividend ratio, computed as $PD_t - A$, into the contribution of expected returns, given by $B_1\hat{\mu}_t$, and the contribution of expected dividend growth rates, given by $B_2\hat{g}_t$. The figure makes clear that, under this model specification, virtually all variation of the price-dividend ratio is driven by expected return variation. The unconditional sample variance decomposition of the price-dividend ratio can readily be

in 2001.

computed from the unconditional sample variances and covariances of our filtered series for expected returns and expected growth rates:

$$\text{Var}\left(PD_t\right) = B_1^2 \text{Var}\left(\hat{\mu}_t\right) + B_2^2 \text{Var}\left(\hat{g}_t\right) + 2B_1 B_2 \text{Cov}\left(\hat{\mu}_t, \hat{g}_t\right), \tag{5.51}$$

where we use the maximum-likelihood estimates for B_1 and B_2 given above. The sample standard deviation of $\hat{\mu}_t$ equals 5.52%, the sample standard deviation of $\hat{g}_t$ equals 3.15% and the correlation equals 30.25%. As B_1 is more than four times as large as B_2, meaning a factor 16 difference in the squares, effectively all variation of the price-dividend ratio is due to expected return variation. To put it in relative terms, 104.7% of the variation in the price-dividend ratio can attributed to variation in expected returns, only 0.8% to variation in expected growth rates, and -2.75% to their covariance. These results are consistent with the existing literature (Cochrane (2006), Campbell (1991)). We show in the next section that this result is a consequence of only allowing for one frequency in expected growth rates. Once we allow for an additional frequency in expected growth rates, which we find to be persistent, this variance decomposition changes drastically.

5.4.4 The information content of the price-dividend ratio

The previous section shows that almost all variation in price-dividend ratios is caused by variation in expected returns. This then raises the question how informative the filtered series for expected returns and in particular expected dividend growth rates are to predict future returns and dividend growth rates.

We find that (i) the predictability for returns, measured as the R-squared generated by $\mu_t^\star$, equals 15.5% and that (ii) the predictability of dividend growth rates, measured as the R-squared generated by g_t, equals 8.0%. Note that we do not use any instruments to achieve this predictive power. It follows naturally from the restrictions of the present-value model. Hence, we show that the price-dividend ratio also contains information about dividend growth rates. This contrasts the results from standard predictive regressions, which suggest that expected returns are forecastable, but dividend growth rates are not (Cochrane (2006)). We show that once we carefully account for the variation in expected returns via the present-value model, the resulting component is a strong predictor of future dividend growth rates.

5.5 Predictability of dividend growth

5.5.1 Persistence of expected growth rates

The previous section shows that imposing a present-value model allows us to uncover a component of the price-dividend ratio that has strong predictive power for future dividend growth. We find that this component is not very persistent as the estimated autoregressive parameters turns out to be negative. In this section we explore an extension of our model in which we consider more flexible dynamics of expected growth rates. More specifically, we allow for two frequencies for expected dividend growth rates:

$$g_t = \gamma_0 + \hat{g}_{1t} + \hat{g}_{2t}, \tag{5.52}$$

where both $\hat{g}_{1t}$ and $\hat{g}_{2t}$ follow a (near-)AR(1) process ($i = 1, 2$):

$$\hat{g}_{it+1} \quad = \quad \gamma_{it}\hat{g}_{it} + \varepsilon^{g}_{it+1}. \tag{5.53}$$

Like before, we define:

$$\gamma_{it} \quad \equiv \quad \lambda_t \gamma_i. \tag{5.54}$$

Appendix 5.A.2 shows that the price-dividend ratio takes the form:

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_{1t} + B_3\hat{g}_{2t}, \tag{5.55}$$

and provides the coefficients of the price-dividend ratio. Also, Appendix 5.A.2 derives the decompositions of changes in the price-dividend ratio and realized stock returns into shocks to expected returns, both components of expected dividend growth, and unexpected dividend growth. Hence, this model allows for two frequencies in expected growth rates of dividends.

We estimate the model by means of maximum likelihood. The estimation results are provided in Table 5.4 along with the bootstrapped standard errors. First, and foremost, we find a second, much more persistent, component in expected dividend growth rates.[27] The autoregressive parameter equals 0.60 at an annual frequency, which is somewhat lower than the value of 0.775 used in the calibration of Bansal and Yaron (2004). This suggests that formulating a first-order autoregressive process for dividend growth seems too restrictive. It is required to accommodate two frequencies to uncover the persistent predictable component of dividend growth rates. Second, the R-squared values improve substantially. The R-squared

[27] See Bansal, Kiku, and Yaron (2006), Bansal, Gallant, and Tauchen (2007), and Lustig, Verdelhan, and Nieuwerburgh (2007) for formal estimation and testing of the Bansal and Yaron (2004) model.

for returns now equals 17.9% and the R-squared for dividend growth 16.1%. This suggests that dividend growth rates are almost as predictable as future returns.

We further consider the counterfactual that expected growth rates have an autoregressive root of 0.9 on an annual frequency. This value is not reliably different from the point estimate reported in Table 5.4. The degree of return predictability then increases even further as measured by the R-squared. Standard predictive regressions typically fail to detect return predictability in the recent decade because the sharp increase in the price-dividend ratio leads to a sharp decline in predicted returns in the 1990s. As returns in this period remained high, this leads to large prediction errors in this period. However, when we consider a highly persistent expected dividend growth rate, the increase in the price-dividend ratio in the 1990s is partly attributed to an increase in the persistent component of expected growth rates and not only to a decrease in expected returns. As such, returns in the 1990s are much better predicted by this model. An investor that would have adhered this world-view would have outperformed an investor who instead relied on predictive regressions.

Finally, note that Bansal and Yaron (2004) assume that the correlation between innovations to the stochastic volatility process, which drives variation in expected returns, and innovations to the long-run risk process, is zero. As such they find that the price-dividend ratio and expected growth rates are positively correlated, i.e., a high price-dividend ratio predicts high future dividend growth rates. In our framework, we uncover a high positive correlation between expected returns and the persistent component of expected dividend growth rates, both in levels and in innovations. As such a positive shock to expected growth rates is accompanied by a positive shock to expected returns. This expected return shock dominates leading to a drop in the price-dividend ratio. We therefore find a negative correlation between the level of expected growth rates and the level of the price-dividend ratio. If we would control for the expected return, we would indeed find a positive correlation between expected growth rates and the price-dividend ratio, which can easily be seen from the expression for the coefficient B_3, which is positive.

5.5.2 Why does the price-dividend ratio move?

Given the persistence of expected dividend growth rates, the variance decomposition of the price-dividend ratio will differ substantially from our earlier results in Section 5.4.4. We now find that 385% of the variation of the price-dividend ratio can be attributed to expected return variation, 92.5% to the variation in the persistent component of expected growth rates and 3.5% to the transient component of expected growth rates. As the correlation between the persistent component of expected growth rates and expected returns is so high, the covariances contribute substantially to the decomposition of the price-dividend ratio,

totalling −358.3%. The remaining −22.7% can attributed to the covariance between expected returns and the transient component of expected dividend growth rates. This shows that the standard result that all variation in the price-dividend ratio can be attributed to expected returns, substantially alters when we allow for a persistent component in expected dividend growth rates. The decomposition of the price-dividend ratio in these components is displayed in Figure 5.9.

5.5.3 Why do prices move?

In Table 5.5 we reconsider the variance decomposition of unexpected returns (compare Table 5.3 for the case where expected dividend growth rates only have a transient component). Equation (5.91) decomposes unexpected stock returns into seven components, four first-order effects and three second-order effects. In line with the decomposition for the price-dividend ratio, we find that shocks to the persistent component of expected growth rates significantly contributes to the variation in unexpected returns. Also, due to the high correlation between expected returns and the persistent component of expected growth rates, the negative covariance between these two components plays a substantial role in the variance decomposition of unexpected returns. We can conclude that allowing for two frequencies in expected growth rates substantially alters our view of what moves the price-dividend ratio and returns. Finally, as discussed in the introduction, we decompose unexpected historical stock returns into shocks to expected returns, the two components of expected dividend growth rates, and shocks to unexpected dividend growth rate. The results are presented in Figure 5.8. The main difference with Figure 5.4 is that shocks in expected returns are now accompanied by offsetting shocks to the persistent component of expected dividend growth rates.

5.5.4 Hypothesis testing within present-value models

Our likelihood-based estimation approach facilitates straightforward hypothesis testing using the likelihood-ratio test. Denote the log-likelihood that corresponds to the unconstrained model by $\mathcal{L}^1$. The log-likelihood that follows from estimating the model under the null hypothesis is denoted by $\mathcal{L}^0$. The likelihood-ratio statistic is then given by:

$$LR = 2(\mathcal{L}^1 - \mathcal{L}^0), \tag{5.56}$$

which is asymptotically chi-squared distributed with the degrees of freedom equal to the number of constrained parameters.

As an application of this test, we formally test whether expected dividend growth contains

a second, more persistent, frequency. The null hypothesis that corresponds to this test reads:

$$H_0 : \gamma_1 = \sigma_{g_1} = \rho_{\mu g_1} = 0, \tag{5.57}$$

and the LR statistic follows a χ_3-distribution. The corresponding p-value equals $p = 0.087$. Hence, we can reject the null hypothesis at the 10%-significance level. Furthermore, the null-hypothesis that expected dividend growth is constant, can be rejected at the 5% significance level.

We also performed a test for the presence of return predictability. That is, we considered the null hypothesis that states that $\delta_1 = \sigma_\mu = \rho_{g\mu} = \sigma_{D\mu} = 0$. Consistent with the results in Table 5.2, which reports small standard errors around δ_1, σ_μ, and $\rho_{g\mu}$, we can strongly reject this hypothesis.

5.6 Extensions

We extend our approach in this section in several directions. In Section 6.1, we decompose the total expected return into the short rate and the equity risk premium. Next, Section 6.2 illustrate how to incorporate information contained in other predictors to further improve the forecasts of future returns and dividend growth rates.

5.6.1 Stochastic short rates

We now consider an extension of our model in which the expected return is decomposed into the short rate and the equity risk premium:

$$y_t = \theta_0 + \hat{y}_t, \tag{5.58}$$

$$\mu_t = \delta_0 + \hat{y}_t + \hat{\mu}_t, \tag{5.59}$$

in which y_t denotes the 1-year short rate and $\hat{\mu}$ the (de-meaned) equity risk premium. The unconditional average of the equity risk premium is given by $\delta_0 - \theta_0$. The dynamics of $\hat{y}_t$ reads:

$$\hat{y}_{t+1} = \lambda_t \theta_1 \hat{y}_t + \epsilon^y_{t+1}, \tag{5.60}$$

and the dynamics of the other variables remains unchanged. We focus on the model in Section 2, with one frequency in expected growth rates, for ease of exposition. Extensions that feature multiple frequencies in either the equity risk premium or expected growth rates

are straightforward to derive. We denote the covariance of innovations to the short rate with innovations to dividend growth by σ_{yD}.

The price-dividend ratio is now of the form:

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t + B_3\hat{y}_t, \tag{5.61}$$

where the coefficients are given by:

$$\begin{aligned}
A &= \left(1 + \frac{\alpha - \alpha^2\delta_1 - \alpha^2\theta_1 + \alpha^3\delta_1\theta_1}{\alpha\delta_1 + \alpha\theta_1 - \sigma_{yD} - \sigma_{D\mu} + \alpha\delta_1\sigma_{yD} + \alpha\theta_1\sigma_{D\mu} - \alpha^2\delta_1\theta_1 - 1}\right)^{-1}, && (5.62)\\
B_1 &= \frac{-A\left(1-\alpha\theta_1\right)}{1 - \alpha\delta_1\left(1+\sigma_{yD}\right) - \alpha\theta_1\left(1+\sigma_{D\mu}\right) + \sigma_{yD} + \sigma_{D\mu} + \alpha^2\delta_1\theta_1}, && (5.63)\\
B_3 &= \frac{-A\left(1-\alpha\delta_1\right)}{1 - \alpha\delta_1\left(1+\sigma_{yD}\right) - \alpha\theta_1\left(1+\sigma_{D\mu}\right) + \sigma_{yD} + \sigma_{D\mu} + \alpha^2\delta_1\theta_1}, && (5.64)\\
B_2 &= \frac{A + B_1\sigma_{D\mu} + B_3\sigma_{rD}}{1 - \gamma_1\alpha}. && (5.65)
\end{aligned}$$

In this model, the shocks to the price-dividend ratio can be attributed to shocks to expected returns, expected growth rates, and short rates. The model can be estimated in the same way as discussed in Section 5.3. All that changes is that we add one measurement equation for the short rates and that we use information on short rates in estimation. We find very similar results, which are available upon request.

5.6.2 Including other predictors

So far, we introduced an efficient method to extract information on expected returns and expected dividend growth rates from price-dividend ratios and dividend growth rates only. We now show how to extend our model to also incorporate information contained in other predictor variables that have been suggested in the literature, see Lamont (1998), Baker and Wurgler (2000), Lettau and Ludvigson (2001), Lettau and Ludvigson (2005), Menzly, Santos, and Veronesi (2004), Lustig and Nieuwerburgh (2005), Piazzesi, Schneider, and Tuzel (2004), and Polk, Thompson, and Vuolteenaho (2006), among others. We wish to extract the information of N predictor variables x_t, $x_t = (x_{1t}, \ldots, x_{Nt})'$, that are relevant to predict future returns and dividend growth rates. To this end, we follow Pástor and Stambaugh (2006) and formulate our model as a predictive system. We add to the transition equation (5.46) and two measurement equations (5.47) and (5.48) the following set of N measurement equations:

$$x_{i,t+1} = \alpha_{0i} + \alpha_{1i}x_{it} + \varepsilon^x_{i,t+1}, \tag{5.66}$$

with $i = 1, \ldots, N$ a set of N additional predictor variables and denote $x_t = (x_{1t}, \ldots, x_{Nt})'$. We further consider the covariance structure of the innovations:

$$\mathrm{Var}_t \begin{pmatrix} \varepsilon^g_{t+1} \\ \varepsilon^\mu_{t+1} \\ \varepsilon^D_{t+1} \\ \varepsilon^x_{t+1} \end{pmatrix} = \begin{pmatrix} \sigma_g^2 & \sigma_{g\mu} & 0 & \sigma'_{gx} \\ \sigma_{g\mu} & \sigma_\mu^2 & \sigma_{D\mu} & \sigma'_{\mu x} \\ 0 & \sigma_{D\mu} & \sigma_D^2 & \sigma'_{xD} \\ \sigma_{gx} & \sigma_{\mu x} & \sigma_{xD} & \Sigma_x \end{pmatrix}, \tag{5.67}$$

with $\varepsilon^x_{t+1} = (\varepsilon^x_{1,t+1}, \ldots, \varepsilon^x_{N,t+1})'$. Pástor and Stambaugh (2006) show that it is the correlation between the innovations in the other predictor variables and the other processes that allows us to incorporate any information useful in capturing the dynamics of expected returns and growth rates. The main advantage of using predictive systems is that we do not a priori impose that expected returns or expected growth rates are affine functions of a set of predictor variables. Instead, we extract any information that is useful in filtering expected returns and growth rates via the correlation in innovations. This model can interpreted as an extension of Pástor and Stambaugh (2006) in which we account for the present-value relationship in estimating the predictive regression for not only returns, but also for dividend growth. Note, however, that including other instruments makes our approach susceptible to the selection bias as described in Ferson, Sarkissian, and Simin (2003). We therefore abstract from taking a stance on which instruments should be included, particularly given the strong predictability in returns and dividend growth rates that we find without specifying such instruments.

5.7 Conclusion

We use a closed-form present-value model to estimate the time series of expected returns and expected dividend growth rates. By imposing this economic model, we show that the aggregate price-dividend ratio has strong predictive power for both future stock market returns and future dividend growth, with R-squared values of 18% and 16%, respectively. We do not rely on VAR-models with prespecified instruments to achieve this result. All the information we use is contained in dividend growth rates and the price-dividend ratio. The resulting time series of expected dividend growth rates turns out to have both a transient and a persistent component.

To obtain these results, we develop a closed-form present-value model that features time-varying expected returns, time-varying expected dividend growth rates, and potentially stochastic interest rates. We develop a likelihood-based estimation framework to empirically analyze the implications of our model. We show that using a present-value model resolves the errors-in-variables problem and greatly alleviates the bias and efficiency problems that have been widely discussed in the predictability literature.

In future work, we consider the case with stochastic volatilities and correlations to study the risk-return tradeoff within a present-value model, which builds upon recent work by Pástor, Sinha, and Swaminathan (2007). Also, our model can be applied to study the time-variation in the cross-section of stock returns, as well as the variation in investment opportunities across countries.

5.A Derivations present-value model

5.A.1 Benchmark model

In this appendix we derive the coefficients of the present-value model in Section 5.2. We start from the present-value relation:

$$PD_t = 1 + \frac{E_t(PD_{t+1}D_{t+1}/D_t)}{1+\mu_t^\star}, \tag{5.68}$$

which we rewrite to:

$$\begin{aligned} PD_t &= 1 + \frac{1+g_t}{1+\mu_t^\star} E_t(PD_{t+1}(1+\varepsilon_{t+1}^D)) \\ &= 1 + (\alpha + \hat{g}_t - \hat{\mu}_t) E_t(PD_{t+1}(1+\varepsilon_{t+1}^D)). \end{aligned} \tag{5.69}$$

We conjecture that the price-dividend ratio is affine in $\hat{\mu}_t$ and $\hat{g}_t$:

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t. \tag{5.70}$$

Substitution of (5.70) into (5.83) leads to:

$$\begin{aligned} A + B_1\hat{\mu}_t + B_2\hat{g}_t &= 1 + (\alpha + \hat{g}_t - \hat{\mu}_t) E_t((A + B_1\hat{\mu}_{t+1} + B_2\hat{g}_{t+1})(1+\varepsilon_{t+1}^D)) \\ &= 1 + (\alpha + \hat{g}_t - \hat{\mu}_t)(A + B_1\sigma_{\mu D}) + \alpha B_1\delta_1\hat{\mu}_t + \alpha B_2\gamma_1\hat{g}_t. \end{aligned} \tag{5.71}$$

We match the coefficients on the constant, $\hat{\mu}_t$, and $\hat{g}_t$:

$$A = 1 + \alpha(A + B_1\sigma_{\mu D}), \tag{5.72}$$

$$B_1 = -(A + B_1\sigma_{\mu D}) + \alpha B_1\delta_1, \tag{5.73}$$

$$B_2 = (A + B_1\sigma_{\mu D}) + \alpha B_2\gamma_1. \tag{5.74}$$

These equations directly result in (5.15)-(5.17).

5.A.2 Two frequencies for expected growth rates

We decompose expected growth rates and expected returns as:

$$g_t = \gamma_0 + \hat{g}_{1t} + \hat{g}_{2t}, \tag{5.75}$$

$$\mu_t = \delta_0 + \hat{\mu}_t, \tag{5.76}$$

where $\hat{g}_{1t}$, $\hat{g}_{2t}$ and $\hat{\mu}_t$ follow a (near-)AR(1) process:

$$\begin{aligned}
\hat{g}_{1t+1} &= \gamma_{1t}\hat{g}_{1t} + \varepsilon^g_{1t+1}, && (5.77)\\
\hat{g}_{2t+1} &= \gamma_{2t}\hat{g}_{2t} + \varepsilon^g_{2t+1}, && (5.78)\\
\hat{\mu}_{t+1} &= \delta_{1t}\hat{\mu}_t + \varepsilon^\mu_{t+1}. && (5.79)
\end{aligned}$$

Like before, we define:

$$\begin{aligned}
\gamma_{1t} &= \lambda_t\gamma_1, && (5.80)\\
\gamma_{2t} &= \lambda_t\gamma_2, && (5.81)\\
\delta_{1t} &= \lambda_t\delta_1, && (5.82)
\end{aligned}$$

where

$$\lambda_t = \frac{\alpha}{\alpha + \hat{g}_{1t} + \hat{g}_{2t} - \hat{\mu}_t}.$$

Recall the present-value relation:

$$PD_t = 1 + \frac{E_t(PD_{t+1}D_{t+1}/D_t)}{1+\mu^\star_t},$$

which we rewrite to:

$$\begin{aligned}
PD_t &= 1 + \frac{1+g_t}{1+\mu^\star_t}E_t(PD_{t+1}(1+\varepsilon^D_{t+1}))\\
&= 1 + (\alpha + \hat{g}_{1t} + \hat{g}_{2t} - \hat{\mu}_t)E_t(PD_{t+1}(1+\varepsilon^D_{t+1})). && (5.83)
\end{aligned}$$

We conjecture that the price-dividend ratio is affine in $\hat{\mu}_t$ and $\hat{g}_t$:

$$PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_{1t} + B_3\hat{g}_{2t}.$$

Substitution of (5.70) into (5.83) leads to:

$$\begin{aligned}
A + B_1\hat{\mu}_t + B_2\hat{g}_{1t} + B_3\hat{g}_{2t} &= 1 + (\alpha + \hat{g}_{1t} + \hat{g}_{2t} - \hat{\mu}_t)E_t((A + B_1\hat{\mu}_{t+1}\\
&\quad + B_2\hat{g}_{1t+1} + B_3\hat{g}_{2t+1})(1+\varepsilon^D_{t+1}))\\
&= 1 + (\alpha + \hat{g}_{1t} + \hat{g}_{2t} - \hat{\mu}_t)(A + B_1\sigma_{\mu D}) + \alpha B_1\delta_1\hat{\mu}_t\\
&\quad + \alpha B_2\gamma_1\hat{g}_{1t} + \alpha B_3\gamma_2\hat{g}_{2t}. && (5.84)
\end{aligned}$$

We match the coefficients on the constant, $\hat{\mu}_t$, and $\hat{g}_t$:

$$\begin{aligned} A &= 1+\alpha(A+B_1\sigma_{\mu D}), && (5.85)\\ B_1 &= -(A+B_1\sigma_{\mu D})+\alpha B_1\delta_1, && (5.86)\\ B_2 &= (A+B_1\sigma_{\mu D})+\alpha B_2\gamma_1, && (5.87)\\ B_3 &= (A+B_1\sigma_{\mu D})+\alpha B_3\gamma_2. && (5.88) \end{aligned}$$

The expected change of the price-dividend ratio is given by:

$$E_t\left[PD_{t+1}\right]-PD_t = B_1\left(\delta_{1t}-1\right)\hat{\mu}_t + B_2\left(\gamma_{1t}-1\right)\hat{g}_{1t} + B_3\left(\gamma_{2t}-1\right)\hat{g}_{2t}, \tag{5.89}$$

and the unexpected change of the price-dividend ratio reads:

$$PD_{t+1}-E_t\left(PD_{t+1}\right) = B_1\varepsilon^{\mu}_{t+1} + B_2\varepsilon^{g}_{1t+1} + B_3\varepsilon^{g}_{2t+1}. \tag{5.90}$$

Similarly, we can decompose the realized stock return into the conditional expected stock return, $\mu^{\star}_t \equiv E_t\left[R_{t+1}\right]$, and the unexpected stock return given by:

$$R_{t+1} = 1+\mu^{\star}_t + h_t\begin{bmatrix} B_1\varepsilon^{\mu}_{t+1}+B_2\varepsilon^{g}_{1,t+1}+B_3\varepsilon^{g}_{2,t+1}+E_t\left(PD_{t+1}\right)\varepsilon^{D}_{t+1} \\ +B_1\left(\varepsilon^{D}_{t+1}\varepsilon^{\mu}_{t+1}-\sigma_{D\mu}\right)+B_2\varepsilon^{D}_{t+1}\varepsilon^{g}_{1,t+1}+B_3\varepsilon^{D}_{t+1}\varepsilon^{g}_{2,t+1} \end{bmatrix}, \tag{5.91}$$

where:

$$h_t \equiv \frac{(1+g_t)}{PD_t-1}. \tag{5.92}$$

The covariance structure of the innovations is given by:

$$\begin{pmatrix} \varepsilon^{g}_{1,t+1} \\ \varepsilon^{g}_{2,t+1} \\ \varepsilon^{\mu}_{t+1} \\ \varepsilon^{D}_{t+1} \end{pmatrix} \sim N\left(0_{3\times 1}, \begin{pmatrix} \sigma^2_{g_1} & 0 & \sigma_{\mu g_1} & 0 \\ 0 & \sigma^2_{g_2} & \sigma_{\mu g_2} & 0 \\ \sigma_{\mu g_1} & \sigma_{\mu g_2} & \sigma^2_{\mu} & \sigma_{D\mu} \\ 0 & 0 & \sigma_{D\mu} & \sigma^2_{D} \end{pmatrix}\right). \tag{5.93}$$

5.B Non-linear filters

In this appendix, we summarize the non-linear filtering techniques we use to recover the time series of expected returns and expected dividend growth rates.

5.B.1 Unscented Kalman filter

The unscented Kalman filter is second-order accurate in non-linear models and was first introduced by Julier and Uhlmann (1997). We follow the notation of Wan and van der Merwe (2001). To initialize the filter, we simulate N trajectories of $\hat{\mu}$ and $\hat{g}$ to compute the unconditional mean ($\hat{g}_0$) and variance (P_0) of $\hat{g}$.

For each observation, we compute the time-t likelihood contribution in the following way:

- Define the augmented state vector $\hat{x}_{t-1} = \left(\hat{g}_{t-1|t-1}, 0, 0, 0\right)'$, with $\hat{g}_{t-1|t-1}$ denoting the filtered value of $\hat{g}_{t-1}$ given the information at time $t-1$. The augmented state vector includes the filtered state and the innovations of the model. For $t = 1$, we have $\hat{x}_0 = (\hat{g}_0, 0, 0, 0)'$. The covariance matrix is given by:

$$P^a_{t-1} = \begin{pmatrix} P_{t-1} & 0 & 0 & 0 \\ 0 & \sigma_g^2 & 0 & \sigma_{\mu g} \\ 0 & 0 & \sigma_D^2 & \sigma_{\mu D} \\ 0 & \sigma_{\mu g} & \sigma_{\mu D} & \sigma_\mu^2 \end{pmatrix}. \tag{5.94}$$

- Compute the so-called sigma points:

$$\chi_{t-1} = \left[\begin{array}{cccc} \hat{x}_{t-1|t-1} & \hat{x}_{t-1|t-1} \times \iota_{1\times L} + \gamma\sqrt{P^a_{t-1}} & \hat{x}_{t-1|t-1} \times \iota_{1\times L} - \gamma\sqrt{P^a_{t-1}} \end{array} \right], \tag{5.95}$$

  where $\gamma = \sqrt{\alpha^2 L}$ and L equals the dimension of the augmented state (4 in our case). The vector $\iota_{1\times L}$ denotes a (row) vector of ones and $\sqrt{A}$ refers to the Cholesky decomposition of the matrix A. The matrix χ_{t-1} therefore has dimensions $L \times (2L+1)$.

- Prediction step:

$$\begin{aligned} \hat{g}^\chi_{t|t-1} &= \left[\frac{\gamma_1 \alpha}{\alpha + \chi_{t-1,(1,:)} - [PD_{t-1} - A - B_2 \chi_{t-1,(1,:)}] B_1^{-1}} \right] \chi_{t-1,(1,:)} + \chi_{(2,:)}, && (5.96) \\ \hat{g}_{t|t-1} &= \hat{g}^\chi_{t|t-1} W_m, && (5.97) \\ P_{t|t-1} &= \sum_{i=0}^{2L} W_i^{(c)} \left(\hat{g}^\chi_{t|t-1(i)} - \hat{g}_{t|t-1} \right)^2, && (5.98) \end{aligned}$$

  where $\alpha \in [1e-4, 1]$. Note that $\hat{g}^\chi_{t|t-1}$ denotes a $(2L+1)$-dimensional row vector. The weights are defined as:

$$\begin{aligned} W_0^{(m)} &= 1 - \alpha^{-2}, && (5.99) \\ W_0^{(c)} &= W_0^{(m)} + \left(3 - \alpha^2\right), && (5.100) \\ W_i^{(m)} &= W_i^{(c)} = \left(2\alpha^2 L\right)^{-1}, \; i = 1, \ldots, 2L. && (5.101) \end{aligned}$$

The prediction grid for dividend growth and the price-dividend ratio are given by:

$$\Delta D^{\chi}_{t|t-1} = (1+\chi_{t-1,(1,:)})(1+\chi_{t-1,(3,:)}), \tag{5.102}$$

$$\begin{aligned} PD^{\chi}_{t|t-1} &= A + \left[\frac{\delta_1 \alpha}{\alpha + \chi_{t-1,(1,:)} - [PD_{t-1} - A - B_2\chi_{t-1,(1,:)}]B_1^{-1}}\right]\left[PD_{t-1} - A - B_2\chi_{t-1,(1,:)}\right] \\ &\quad + B_2\left[\frac{\gamma_1 \alpha}{\alpha + \chi_{t-1,(1,:)} - [PD_{t-1} - A - B_2\chi_{t-1,(1,:)}]B_1^{-1}}\right]\chi_{t-1,(1,:)} \\ &\quad + B_1\chi_{t-1,(4,:)} + B_2\chi_{t-1,(2,:)}, \end{aligned} \tag{5.103}$$

and the predictions read:

$$\Delta D_{t|t-1} = \Delta D^{\chi}_{t|t-1} W^{(m)}, \tag{5.104}$$

$$PD_{t|t-1} = PD^{\chi}_{t|t-1} W^{(m)}. \tag{5.105}$$

- Define the prediction errors in the observables:

$$\zeta_i \equiv \begin{pmatrix} \Delta D^{\chi}_{t|t-1(i)} - \Delta D_{t|t-1} \\ PD^{\chi}_{t|t-1(i)} - PD_{t|t-1} \end{pmatrix}. \tag{5.106}$$

The covariance matrix of the prediction errors of the observables is in turn given by:

$$\Sigma_{\zeta\zeta,t} = \sum_{i=0}^{2L} W_i^{(c)} \zeta_i \zeta_i', \tag{5.107}$$

and the covariance matrix of the prediction errors of the observables and the latent process $\hat{g}$ by:

$$\Sigma_{\hat{g}\zeta,t} = \sum_{i=0}^{2L} W_i^{(c)} \left(\hat{g}^{\chi}_{t|t-1(i)} - \hat{g}_{t|t-1}\right) \zeta_i'. \tag{5.108}$$

- Update step:
The Kalman gain can then be defined as:

$$K_t = \Sigma_{\hat{g}\zeta,t}\Sigma^{-1}_{\zeta\zeta,t}. \tag{5.109}$$

The latent state is subsequently updated using the new observations:

$$\hat{g}_{t|t} = \hat{g}_{t|t-1} + K_t \begin{pmatrix} \Delta D_t - \Delta D_{t|t-1} \\ PD_t - PD_{t|t-1} \end{pmatrix}, \tag{5.110}$$

and the covariance matrix of $\hat{g}$:

$$P_t = P_{t-1} - K_t \Sigma_{\zeta\zeta,t} K_t'. \tag{5.111}$$

The likelihood can then be constructed in the same way as in case of a standard Kalman filter.

5.B.2 Particle filter

We also employ the particle filter to solve the filtering problem to analyze the accuracy of the unscented Kalman filter. The particle filter has proved to be useful in economics (see Fernández-Villaverde and Rubio-Ramírez (2006)) and allows to account for the inherent non-linearities of the present-value model. Section 5.B.3 provides the theoretical background, while 5.B.4 provides further practical details on implementation.

5.B.3 Theoretical background

The inference problem is characterized by one transition equation as in (5.46):

$$\hat{g}_{t+1} = \gamma_{1t} \hat{g}_t + \varepsilon^g_{t+1}, \tag{5.112}$$

and two measurement equations:

$$\begin{aligned} \frac{D_{t+1}}{D_t} &= 1 + g_t + (1 + g_t)\varepsilon^D_{t+1}, & (5.113)\\ PD_{t+1} &= A + \delta_{1t}(PD_t - A - B_2 \hat{g}_t) + B_2 \gamma_{1t} \hat{g}_t + B_1 \varepsilon^\mu_{t+1} + B_2 \varepsilon^g_{t+1} \\ &\equiv \mu^{PD}_t + B_1 \varepsilon^\mu_{t+1} + B_2 \varepsilon^g_{t+1}, & (5.114) \end{aligned}$$

where the conditional expectation of PD_{t+1}, i.e., μ^{PD}_t, depends only on PD_t and $\hat{g}_t$ using (5.16). We use the time series of dividend growth $\{D_t/D_{t-1}\}_{t=1}^T$ and price-dividend ratios $\{PD_t\}_{t=0}^T$ in estimation and filtering.

The aim of this section is to construct the likelihood of the observed time series to estimate the parameters of the present-value model with maximum likelihood, and to filter $\{\hat{g}_t\}$ in turn. To this end, we put further structure on the innovations of the model. In particular, we assume all innovations to be Gaussian and, for identification of the model, that the correlation between expected and unexpected dividend growth innovations equals

zero. The innovations then satisfy:

$$\begin{pmatrix} \varepsilon_{t+1}^{g} \\ \varepsilon_{t+1}^{\mu} \\ \varepsilon_{t+1}^{D} \end{pmatrix} \sim N\left(0_{3\times 1}, \begin{pmatrix} \sigma_g^2 & \sigma_{\mu g} & 0 \\ \sigma_{\mu g} & \sigma_{\mu}^2 & \sigma_{D\mu} \\ 0 & \sigma_{D\mu} & \sigma_D^2 \end{pmatrix}\right), \tag{5.115}$$

As a result, the distribution of the innovations to $\hat{g}_t$, $\{D_t/D_{t-1}\}$, and $\{PD_t\}$ is given by:

$$\begin{pmatrix} \hat{g}_{t+1} - E_t(\hat{g}_{t+1}) \\ PD_{t+1} - E_t(PD_{t+1}) \\ D_{t+1}/D_t - E_t(D_{t+1}/D_t) \end{pmatrix} \sim N\left(0_{3\times 1}, \begin{pmatrix} \sigma_g^2 & B_1\sigma_{\mu g} & 0 \\ B_1\sigma_{\mu g} & B_1^2\sigma_{\mu}^2 + B_2^2\sigma_g^2 + 2B_1B_2\sigma_{g\mu} & (1+g_t)B_1\sigma_{D\mu} \\ 0 & (1+g_t)B_1\sigma_{D\mu} & (1+g_t)^2\sigma_D^2 \end{pmatrix}\right).$$

The likelihood of the observed series depends in total on eight parameters that we collect in the parameter vector Θ:

$$\Theta = \{\gamma_0, \gamma_1, \delta_0, \delta_1, \sigma_g, \sigma_D, \sigma_{\mu}, \sigma_{g\mu}, \sigma_{D\mu}\}. \tag{5.116}$$

We now factorize the likelihood as:[28]

$$\mathcal{L}\left(y^T; \Theta\right) = \prod_{t=1}^{T} \ell\left(y_t \mid y^{t-1}; \Theta\right), \tag{5.117}$$

with $y_t = (PD_t, D_t/D_{t-1})$, $y^t = \{y_1, \ldots, y_t\}$ and $y^0 = \{PD_0\}$. Subsequently, we can write:

$$\begin{aligned} \mathcal{L}\left(y^T; \Theta\right) &= \prod_{t=1}^{T} \ell\left(y_t \mid y^{t-1}; \Theta\right) && (5.118) \\ &= \prod_{t=1}^{T} \int\int \ell\left(y_t \mid y^{t-1}, \varepsilon_g^t, g_0; \Theta\right) \ell\left(\varepsilon_g^t, g_0 \mid y^{t-1}; \Theta\right) \mathrm{d}\varepsilon_g^t \mathrm{d}g_0, && (5.119) \end{aligned}$$

in which $\varepsilon_g^t = \{\varepsilon_1^g, \ldots, \varepsilon_t^g\}$.

The main complication is that the likelihood in (6.24) cannot be computed analytically and we need to resort to numerical techniques instead. Specifically, we use simulation-based techniques to evaluate the integrals and to construct the likelihood in turn. The main idea is that once we have N draws that originate from the sequence of densities:

$$\left\{\ell\left(\varepsilon_g^t, g_0 \mid y^{t-1}; \Theta\right)\right\}_{t=1}^{T}, \tag{5.120}$$

[28]The remainder of this appendix closely follows Fernández-Villaverde and Rubio-Ramírez (2004) and Fernández-Villaverde and Rubio-Ramírez (2006).

which we denote by:

$$\left\{\left\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right\}_{i=1}^{N}\right\}_{t=1}^{T}, \tag{5.121}$$

then we can use these draws to approximate the likelihood as an application of the law of large numbers:

$$\mathcal{L}\left(y^T;\Theta\right) \simeq \prod_{t=1}^{T} \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_t \mid y^{t-1}, \varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i};\Theta\right). \tag{5.122}$$

The particle filter can be used to efficiently simulate from $\left\{\ell\left(\varepsilon_g^t, g_0 \mid y^{t-1};\Theta\right)\right\}_{t=1}^{T}$ to approximate the likelihood as in (5.122). The crucial step of the particle filter is the updating step, i.e., how to construct N random draws, $\left\{\{\varepsilon_g^{t,i}, g_0^{t,i}\}_{i=1}^{N}\right.$, from $\ell\left(\varepsilon_g^t, g_0 \mid y^t;\Theta\right)$, once we have available N random draws, $\left\{\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right\}_{i=1}^{N}$, from $\ell\left(\varepsilon_g^t, g_0 \mid y^{t-1};\Theta\right)$. The following proposition from Fernández-Villaverde and Rubio-Ramírez (2004) shows exactly how this can be accomplished.

Proposition 5.1. *Let* $\left\{\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right\}_{i=1}^{N}$ *be* N *random draws from* $\ell\left(\varepsilon_g^t, g_0 \mid y^{t-1};\Theta\right)$. *Let in addition* $\left\{\tilde{\varepsilon}_g^i, \tilde{g}_0^i\right\}$ *indicate a draw with replacement from* $\left\{\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right\}_{i=1}^{N}$, *with* q_t^i *the probability of* $\left(\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right)$ *being drawn for all* i, *and* q_t^i *given by:*

$$q_t^i = \frac{\ell\left(y_t \mid y^{t-1}, \varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i};\Theta\right)}{\sum_{i=1}^{N} \ell\left(y_t \mid y^{t-1}, \varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i};\Theta\right)}. \tag{5.123}$$

Then $\left\{\tilde{\varepsilon}_g^i, \tilde{g}_0^i\right\}_{i=1}^{N}$ *is a draw from* $\ell\left(\varepsilon_g^t, g_0 \mid y^t;\Theta\right)$.

Hence, proposition 5.1 shows that given a draw from the distribution of the innovations and initial condition of the transition equation, and conditional on the time-$(t-1)$ information, we can update these draws so that they correspond to a draw from the distribution conditional on the time-t information. This requires us to resample using the q_t^i probabilities. Along these lines, we can construct a sequence of draws (particles), $\left\{\left\{\varepsilon_g^{t|t-1,i}, g_0^{t|t-1,i}\right\}_{i=1}^{N}\right\}_{t=1}^{T}$, from the sequence of distributions $\left\{\ell\left(\varepsilon_g^t, g_0 \mid y^{t-1};\Theta\right)\right\}_{t=1}^{T}$. The resulting series are in turn used to construct the likelihood as in (5.122). To simulate from the unconditional distribution of expected growth rates, we use a sufficiently long burn-in period so that the initial conditions do not impact our estimate of the likelihood anymore.

5.B.4 Practical implementation

In this section, we provide a brief implementation guide to the particle filter for our model. We denote the number of particles by N. We fix N at 60,000, which is large enough to reduce the sampling uncertainty so that it will have no impact on the final results. The main procedure is as follows:

- **Initialization:** Simulate N trajectories of length $T_{\text{burn in}}$ of $(\hat{g}, \hat{\mu})$ starting from $\hat{g} = \hat{\mu} = 0$. This generates the initial distribution of both latent processes. Denote the resulting state variables by $\hat{g}^{0|0,i}$ for the initialized distribution.

- **Prediction:** From each of these trajectories, draw N innovation from $N(0, \sigma_g^2)$, which we indicate by $\varepsilon_1^{g,i}$. We use these N innovations to form the prediction, i.e.:

$$\hat{g}^{1|0,i} = \frac{\alpha}{\alpha + \hat{g}^{0|0,i} - \hat{\mu}^{0|0,i}} \gamma_1 \hat{g}^{0|0,i} + \varepsilon_1^{g,i}. \tag{5.124}$$

- **Filtering:** We now need to incorporate the time-1 information to update our estimate of the latent process $\hat{g}^{1|0,i}$. At time 1, we observe dividend growth D_1/D_0 and the price-dividend ratio PD_1. To this end, we first compute the resampling weights, q_1^i, which correspond to the (normalized) likelihood of $(D_1/D_0, PD_1)$. In particular, these weights are given by:

$$q_t^i = \frac{\ell(D_1/D_0, PD_1 \mid \hat{g}_1 = \hat{g}^{1|0,i}, \varepsilon_1^{g,i})}{\sum_{i=1}^N \ell(D_1/D_0, PD_1 \mid \hat{g}_1 = \hat{g}^{1|0,i}, \varepsilon_1^{g,i})}. \tag{5.125}$$

  This conditional likelihood can be computed easily using (5.113) and (5.114), as $(D_1/D_0, PD_1, \varepsilon_1^g)$, conditional on $\hat{g}_0$, are jointly normal. Intuitively, q_t^i indicates how likely a particular simulated path is, given that we observe D_1/D_0 and PD_1. Very extreme particles that are unlikely to lead to the observed dividend growth and price-dividend ratio receive a small weight. Likewise, more likely particles receive a relatively larger weight.

- **Sampling:** In the sampling step, we actually update the filtered process $\hat{g}^{1|0,i}$ to $\hat{g}^{1|1,i}$. To this end, we draw N times, with replacement and probabilities $\{q_1^1, \ldots, q_1^i, \ldots, q_1^N\}$, from $\{g^{1|0,1}, \ldots, g^{1|0,i}, \ldots, g^{1|0,N}\}$. The resulting draws are then indicated by $g^{1|1,i}$. We average these series to obtain the filtered series, i.e.:

$$\hat{g}^{1|1} = \frac{1}{N} \sum_{i=1}^{N} \hat{g}^{1|1,i}. \tag{5.126}$$

- **Time 2,. . . ,T:** We now iterate the algorithm from the prediction step onwards and continue up to time T.

Finally, we have that by the law of large numbers:

$$\ell(D_t/D_{t-1}, PD_t \mid (D/D_{-1})^{t-1}, PD^{t-1}) \simeq \frac{1}{N}\sum_{i=1}^{N} \ell(D_t/D_{t-1}, PD_t \mid (D/D_{-1})^{t-1}, PD^{t-1}, \hat{g}^{t|t-1,i}, \varepsilon_t^{g,i}),$$

which allows us to construct the likelihood. These steps results in (i) a simulated likelihood and (ii) a filtered time series of $\{\hat{g}_t\}$ and $\{\hat{\mu}_t\}$. We optimize the likelihood over the model parameters. Since simulated likelihoods inevitably not differentiable, Newton-type optimizer do not tend to work very well. Instead, we use simulated annealing to perform the optimization. This simulation-based optimization method is designed specifically for non-monotone functions, see Goffe, Ferrier, and Rogers (1994).

5.C Tables and figures

	True value	Mean	St.dev	5%	25%	50%	75%	95%	RMSE
β_{OLS}	-0.0064	-0.0106	0.0060	-0.0211	-0.0141	-0.0101	-0.0066	-0.0017	**0.00730**
$\beta_{\text{OLS, Adapted}}$	-0.0064	-0.0096	0.0053	-0.0191	-0.0127	-0.0091	-0.0060	-0.0019	**0.00616**
β_{ML}	-0.0064	-0.0088	0.0032	-0.0148	-0.0107	-0.0084	-0.0065	-0.0044	**0.00405**
$\gamma_{0,\text{OLS}}$	0.02500	0.0249	0.0186	-0.0059	0.0123	0.0248	0.0376	0.0551	0.01862
$\gamma_{0,\text{ML}}$	0.02500	0.0248	0.0188	-0.0065	0.0124	0.0249	0.0375	0.0555	0.01882
$\delta_{0,\text{ML}}$	0.0600	0.0619	0.0206	0.0292	0.0485	0.0616	0.0749	0.0945	0.02071
$\delta_{1,\text{ML}}$	0.8500	0.7848	0.1001	0.6072	0.7274	0.7960	0.8514	0.9198	0.11947
$\sigma_{D,\text{ML}}$	0.1400	0.1385	0.0132	0.1172	0.1293	0.1383	0.1474	0.1607	0.01332
$\sigma_{\mu,\text{ML}}$	0.0180	0.0241	0.0090	0.0118	0.0178	0.0229	0.0293	0.0401	0.01086
$\rho_{\mu D,\text{ML}}$	0.0516	-0.0006	0.1329	-0.2187	-0.0919	-0.0008	0.0902	0.2169	0.14279

Table 5.1: Comparison of estimation methods
We simulate from our present-value model with constant growth rates and time-varying discount rates with the following set of parameters that are close to the maximum-likelihood estimates: $\delta_0 = 0.06$, $\delta_1 = 0.85$, $\gamma_0 = 0.025$, $\sigma_D = 0.14$, $\sigma_\mu = 0.018$, $\sigma_{D\mu} = 0.00013$. We then estimate the parameters with the following three methods: (i) OLS regression of returns on the price-dividend ratio as in equation (5.37), (ii) adapted OLS regression of returns on the price-dividend ratio as in equation (5.41), and (iii) maximum-likelihood estimation. For the first two estimation procedures, we recover only the slope coefficient, whereas we can estimate all parameters when we employ maximum-likelihood estimation.

Panel A: Maximum-likelihood estimates

	Estimate	S.e.		Estimate	S.e.
δ_0	0.1163	(0.0187)	γ_0	0.0811	(0.0173)
δ_1	0.8243	(0.0807)	γ_1	-0.3347	(0.4186)
σ_μ	0.0282	(0.0192)	σ_g	0.0692	(0.0360)
$\rho_{D\mu}$	0.0404	(0.1795)	σ_D	0.1030	(0.0177)
$\rho_{g\mu}$	0.7913	(0.2810)			

Panel B: Implied present-value model parameters

A	27.93	α	0.96
B_1	-136.31	B_2	29.10

Panel C: R-squared values

$R^2_{Returns}$	15.5%	R^2_{Div}	8.0%

Table 5.2: Estimation results of the present-value model
We present the estimation results of the present-value model of Section 5.2. The model is estimated by maximum likelihood, see Section 5.3, using data from 1946 to 2005 on the dividend growth rate and the price-dividend ratio. Panel A presents the estimates of the coefficients of the underlying processes (bootstrapped standard errors between parentheses). Panel B reports the resulting coefficients of the present-value model ($PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_t$) and the constant $\alpha = 1 + \gamma_0 - \delta_0$. In Panel C, we report the R-squared values for returns and dividend growth rates.

		(1)	(2)	(3)	(4)	(5)
(1):	$h_t B_1 \varepsilon^\mu_{t+1}$	110.0%	-14.2%	-22.9%	0.2%	0.3%
(2):	$h_t B_2 \varepsilon^g_{t+1}$	-14.2%	2.8%	-5.2%	0.4%	-0.1%
(3):	$h_t E_t(PD_{t+1}) \varepsilon^D_{t+1}$	-22.9%	-5.2%	75.3%	-3.6%	0.3%
(4):	$h_t B_1 (\varepsilon^\mu_{t+1}\varepsilon^D_{t+1} - \sigma_{D\mu})$	0.2%	0.36%	-3.6%	2.2%	-0.2%
(5):	$h_t B_2 (\varepsilon^g_{t+1}\varepsilon^D_{t+1})$	0.3%	-0.1%	0.3%	-0.2%	0.0%

Table 5.3: Variance decomposition of stock returns
We present the variance decomposition of unexpected stock returns. Unexpected stock returns are decomposed into shocks to expected returns (1), shocks to expected dividend growth rates (2), unexpected dividend growth rates (3), and two second-order terms ((4) and (5)), see equation (5.49). We compute the time series for each of the processes using observations on dividend growth, price-dividend ratios, and the filtered series to construct a time series for (1)-(5). We then determine the unconditional covariance matrix.

Panel A: Maximum-likelihood estimates

	Estimate	S.e.		Estimate	S.e.
δ_0	0.1172	(0.0227)	γ_0	0.0824	(0.0215)
δ_1	0.8404	(0.4192)	γ_1	0.5984	(0.1875)
			γ_2	-0.7300	(0.2972)
σ_μ	0.0399	(0.0259)	σ_{g1}	0.0437	(0.0321)
$\rho_{D\mu}$	0.0641	(0.2763)	σ_{g2}	0.0345	(0.0259)
$\rho_{\mu g1}$	0.9496	(0.3988)	σ_D	0.1026	(0.0181)
$\rho_{\mu g2}$	0.3069	(0.3829)			

Panel B: Implied present-value model parameters

A	26.82	α	0.96
B_1	-142.70	B_2	15.85
B_3	59.10		

Panel C: R-squared values

$R^2_{Returns}$	17.9%	R^2_{Div}	16.1%

Table 5.4: Estimation results of the present-value model with two frequencies for expected growth rates

We present the estimation results of the present-value model of Section 5.5. The model is estimated by maximum likelihood, see Section 5.3 using data from 1946 to 2005 on the dividend growth rate and the price-dividend ratio. Panel A presents the estimates of the coefficients of the underlying processes (bootstrapped standard errors between parentheses). Panel B reports the resulting coefficients of the present-value model ($PD_t = A + B_1\hat{\mu}_t + B_2\hat{g}_{1t} + B_3\hat{g}_{2t}$) and the constant $\alpha = 1 + \gamma_0 - \delta_0$. In Panel C we report the R-squared values for returns and dividend growth rates.

		(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1):	$h_t B_1 \varepsilon^{\mu}_{t+1}$	269.9%	-117.9%	0.3%	-31.9%	-4.9%	2.3%	0.0%
(2):	$h_t B_2 \varepsilon^{g}_{1,t+1}$	-117.9%	51.8%	0.2%	10.1%	2.2%	-1.0%	0.0%
(3):	$h_t B_3 \varepsilon^{g}_{2,t+1}$	0.3%	0.2%	0.5%	-5.7%	0.0%	0.0%	0.0%
(4):	$h_t E_t(PD_{t+1}) \varepsilon^{D}_{t+1}$	-31.9%	10.1%	-5.7%	70.6%	-1.6%	0.6%	0.1%
(5):	$h_t B_1 (\varepsilon^{\mu}_{t+1} \varepsilon^{D}_{t+1} - \sigma_{D\mu})$	-4.9%	2.2%	0.0%	-1.6%	4.8%	-2.1%	0.1%
(6):	$h_t B_2 (\varepsilon^{g}_{1,t+1} \varepsilon^{D}_{t+1})$	2.3%	-1.0%	0.0%	0.6%	-2.1%	0.9%	0.0%
(7):	$h_t B_3 (\varepsilon^{g}_{2,t+1} \varepsilon^{D}_{t+1})$	0.0%	0.0%	0.0%	0.1%	0.1%	0.0%	0.0%

Table 5.5: Variance decomposition of stock returns
We present the variance decomposition of unexpected stock returns. Unexpected stock returns are decomposed into shocks to expected returns (1), shocks to the persistent component of expected dividend growth rates (2), shocks to the transient component of expected dividend growth rates (3), unexpected dividend growth rates (4), and three second-order terms ((5), (6), and (7)), see (5.91). We compute the time series for each of the processes using observations on dividend growth, price-dividend ratios, and the filtered series to construct a time series for (1)-(7). We then determine the unconditional covariance matrix.

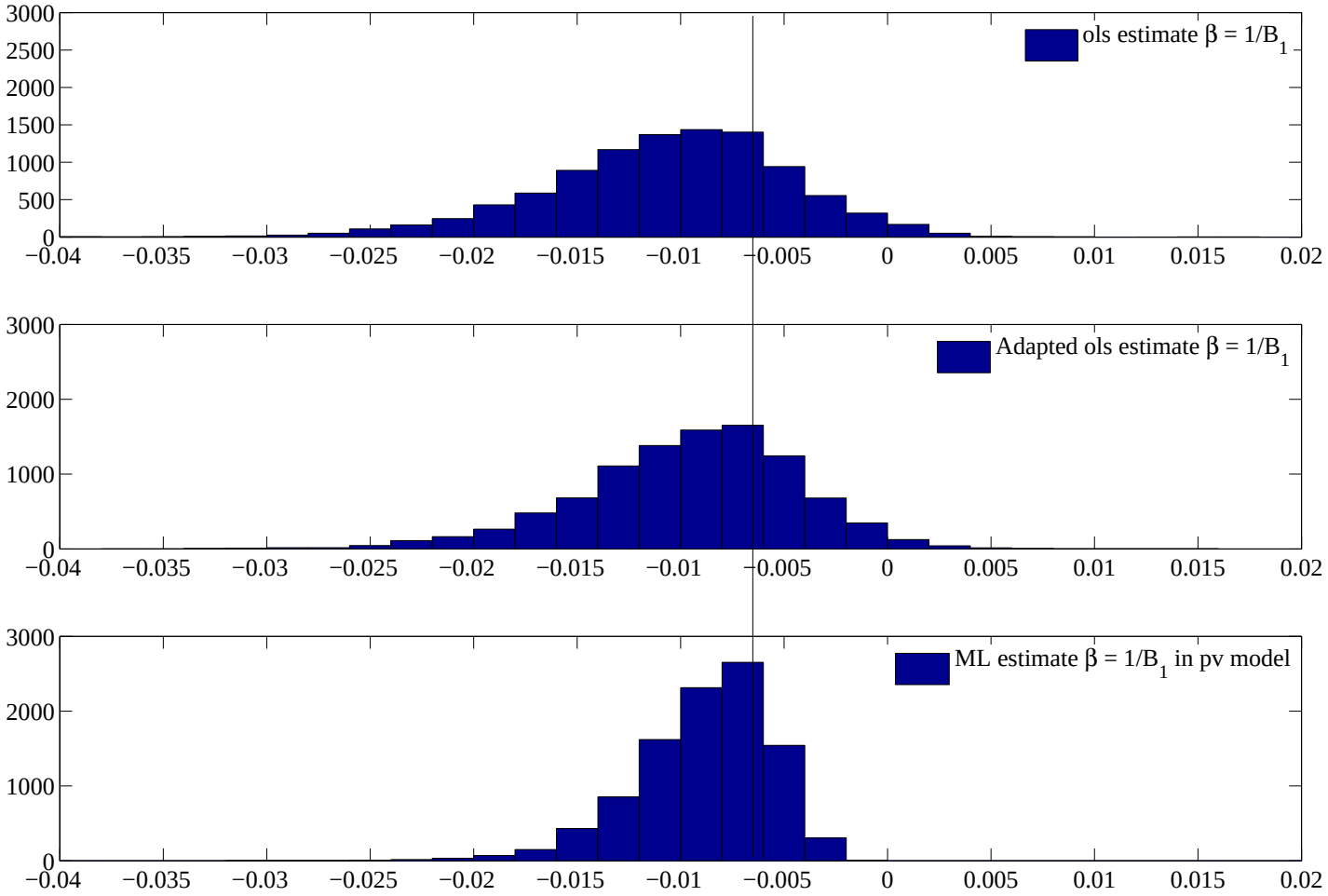


Figure 5.1: Histograms of three estimators
Histograms of three estimators for $\beta = 1/B_1$ in the present-value model with constant growth rates and time-varying discount rates: (i) OLS regression of returns on the price-dividend ratio as in equation (5.37), (ii) adapted OLS regression of returns on the price-dividend ratio as in equation (5.41), and (iii) maximum-likelihood estimation.

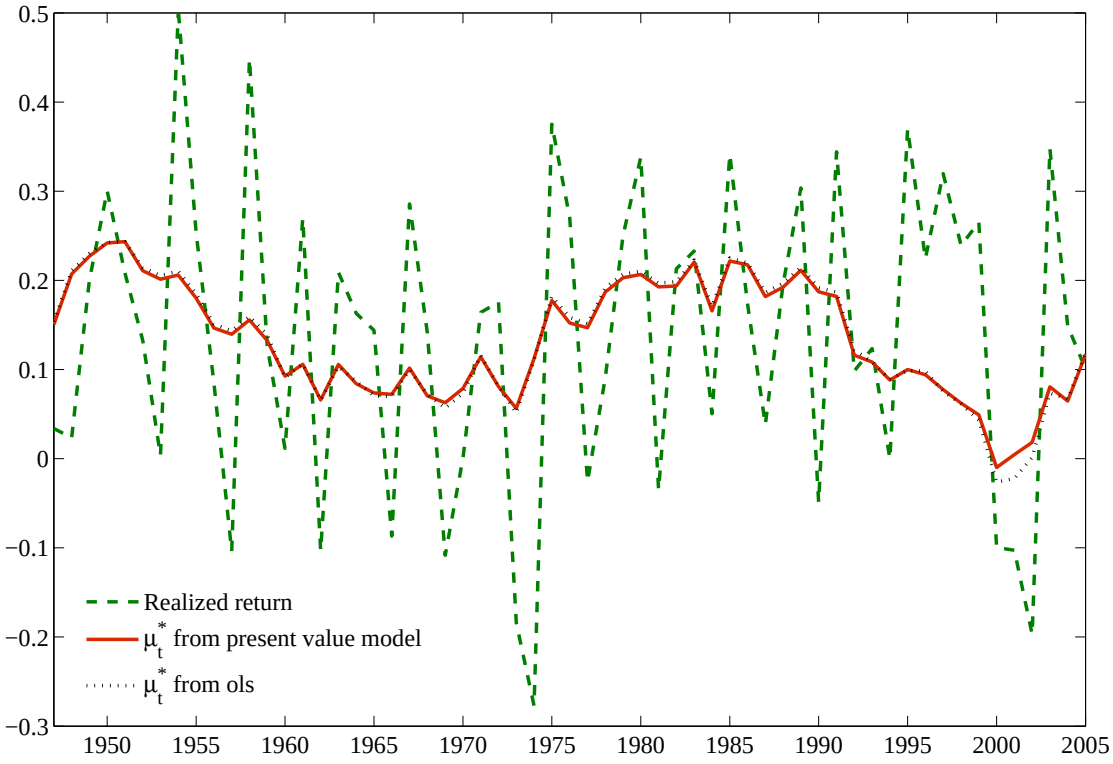


Figure 5.2: Time series of realized and expected returns
We plot the time series of realized and expected returns ($\mu_t^\star$) over the sample 1946-2005. The red solid line corresponds to the filtered expected return series from the present-value model. The black dotted line represents the fitted values from a predictive regression of realized returns on the price-dividend ratio. The green dashed line corresponds to realized returns.

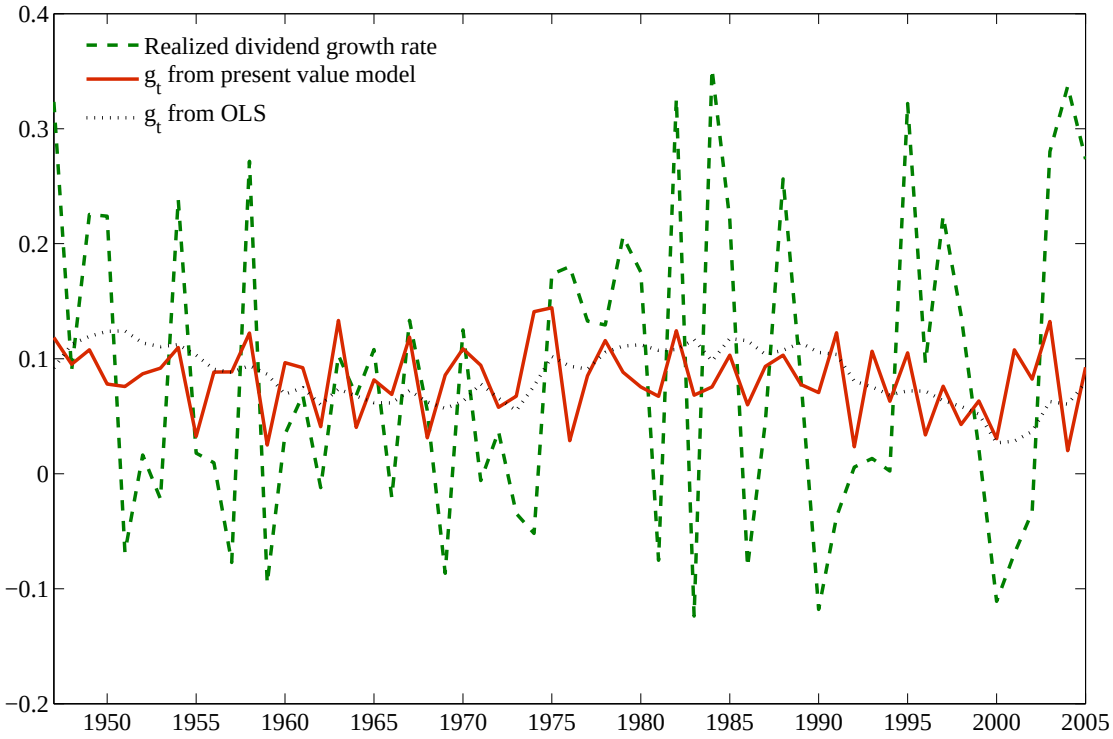


Figure 5.3: Time series of realized and expected dividend growth rates
We plot the time series of realized and expected dividend growth rates (g_t) over the sample 1946-2005. The red solid line corresponds to the filtered series from the present-value model. The black dotted line represents the fitted values from a predictive regression of realized dividend growth rates on the price-dividend ratio. The green dashed line corresponds to realized dividend growth.

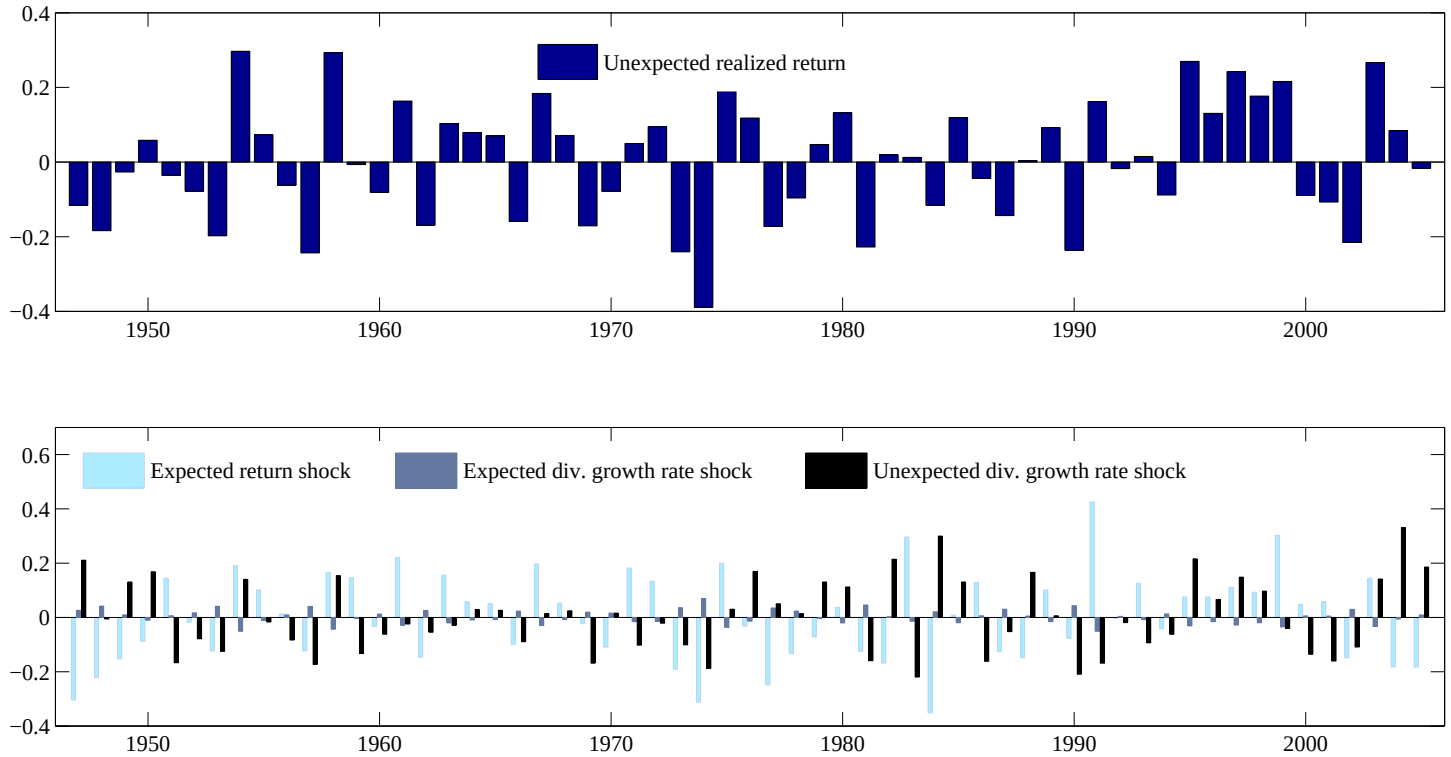


Figure 5.4: Decomposition of unexpected returns: first-order effects
We plot the unexpected return in the top panel. In the bottom panel we plot, in color, the first-order contribution to the unexpected return for the three types of shocks ($\mu^\star$, g and unexpected dividend growth) as given in equation (5.28). The first-order effect of the unexpected return contribution of ε^{μ}_{t+1} is given by $\varepsilon^{\mu}_{t+1} B_1 h_t$ (recall that B_1 is negative), of ε^{g}_{t+1} it is given by $\varepsilon^{g}_{t+1} B_2 h_t$ (recall that B_2 is positive), and of ε^{D}_{t+1} it is given by $\varepsilon^{D}_{t+1} E_t\left(PD_{t+1}\right)$.

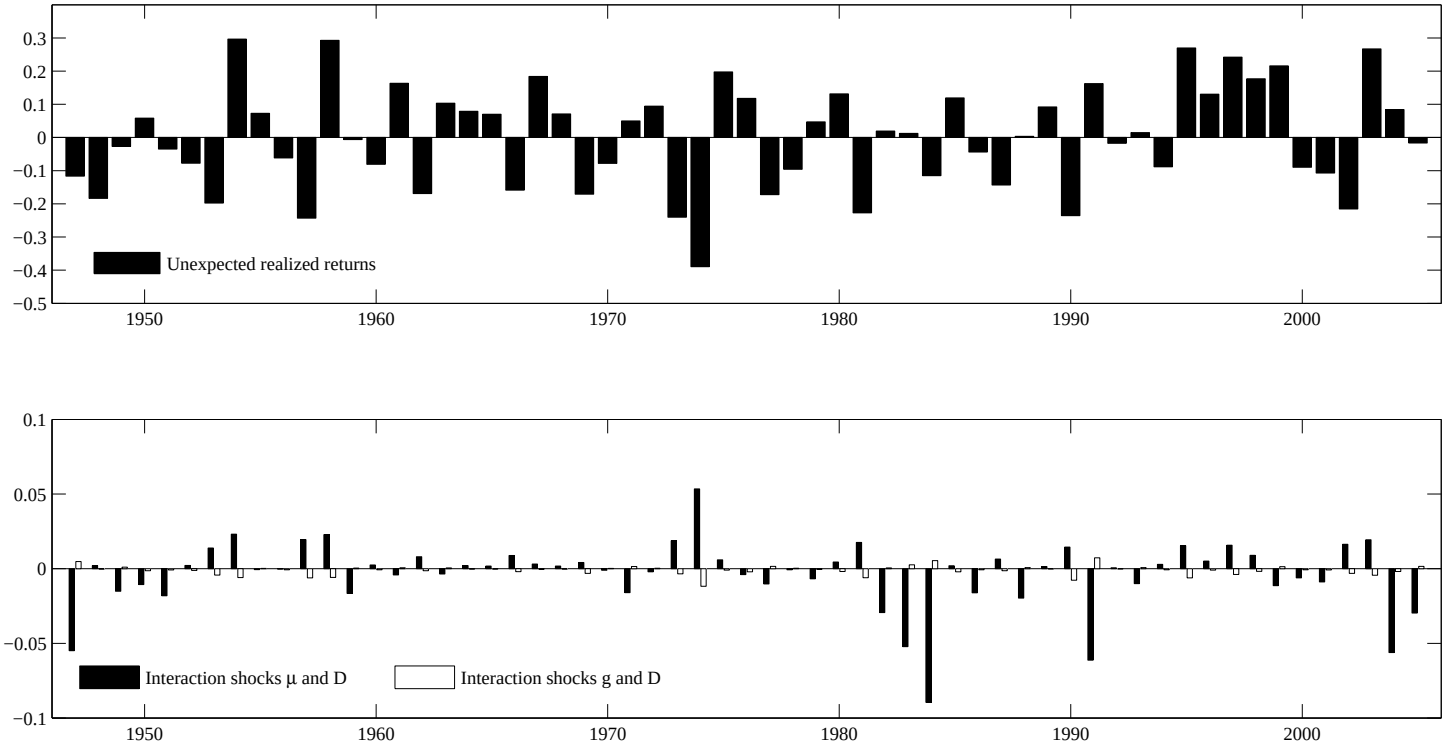


Figure 5.5: Decomposition of unexpected returns: second-order effects
We plot the unexpected return in the top panel. In the bottom panel we plot the second-order contribution to the unexpected return as given in equation (5.28). The second-order effect of the unexpected return contribution of $\varepsilon^{\mu}_{t+1}\varepsilon^{D}_{t+1}$ is given by $\varepsilon^{\mu}_{t+1}\varepsilon^{D}_{t+1}B_1 h_t$ and of $\varepsilon^{g}_{t+1}\varepsilon^{D}_{t+1}$ it is given by $\varepsilon^{g}_{t+1}\varepsilon^{D}_{t+1}B_2 h_t$

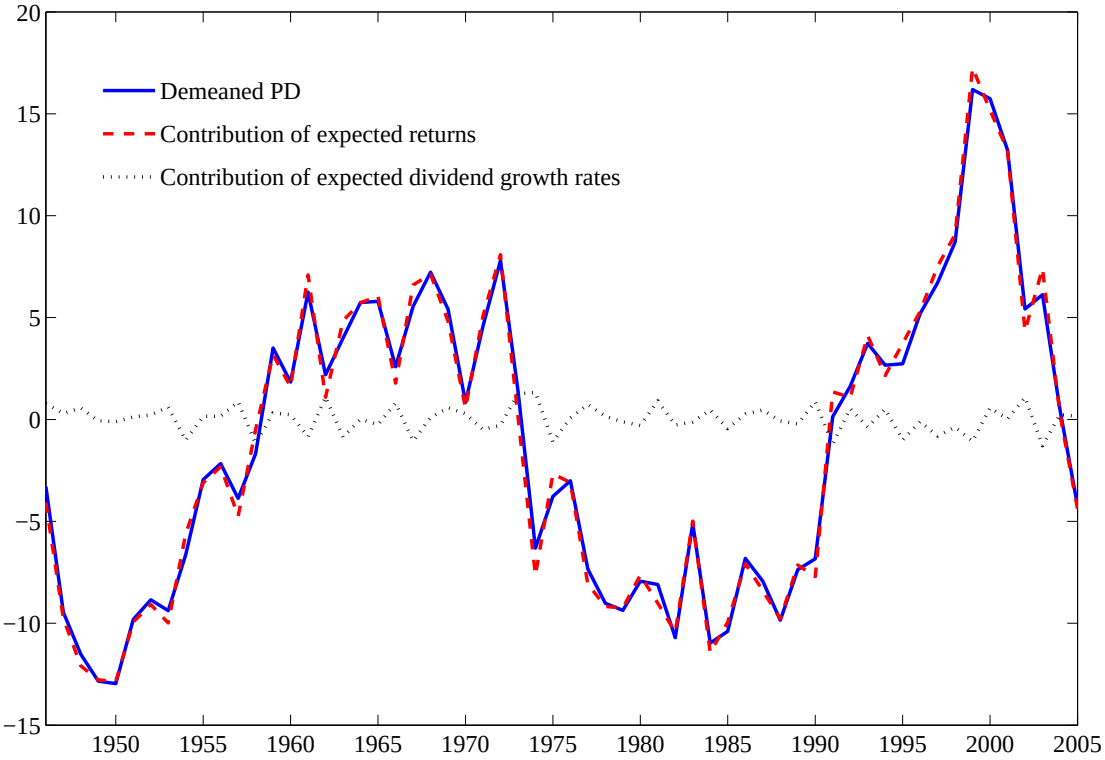


Figure 5.6: Price-dividend decomposition
We decompose the demeaned price-dividend ratio, given by $(PD_t - A)$, into the contribution of expected returns, given by $B_1\hat{\mu}_t$, and the contribution of expected dividend growth rates given by $B_2\hat{g}_t$.

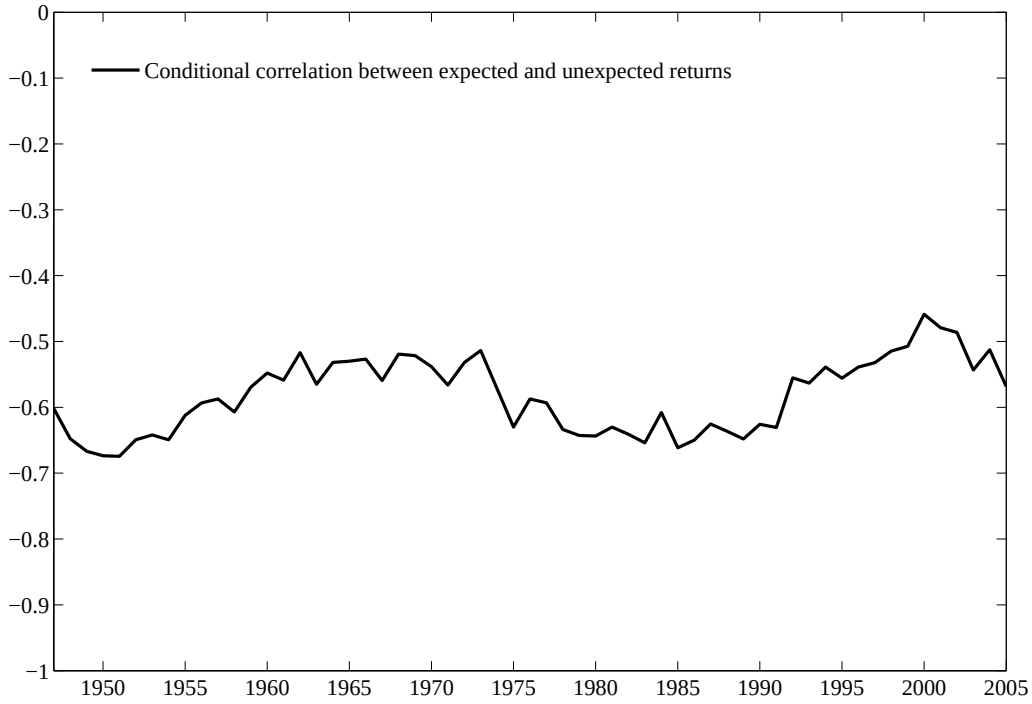


Figure 5.7: Conditional correlation between expected and unexpected returns
Conditional correlation between expected and unexpected returns computed as the ratio of expression (5.31) and the conditional standard deviations of expected returns (σ_μ) and realized returns (see equation (5.30))

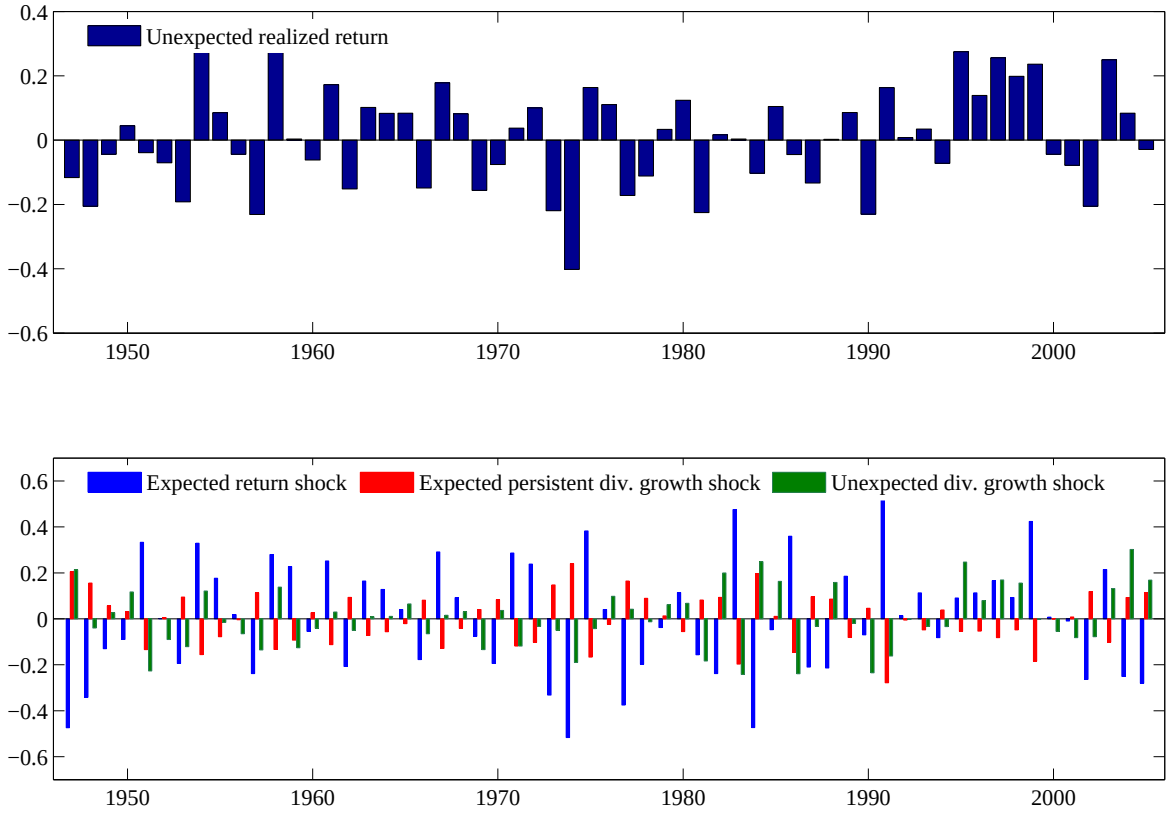


Figure 5.8: Decomposition of unexpected returns: first-order effects
We plot the unexpected return in the top panel. In the bottom panel we plot, in color, the first-order contribution to the unexpected return for the three types of shocks ($\mu^\star$, g_2 and unexpected dividend growth) as given in equation (5.28). We omit the shocks to g_1 for expositional reasons as they are negligible due to the low persistence of this component of expected growth rates. The first-order effect of the unexpected return contribution of ε^{μ}_{t+1} is given by $\varepsilon^{\mu}_{t+1} B_1 h_t$ (recall that B_1 is negative), of ε^{g}_{t+1} it is given by $\varepsilon^{g}_{2,t+1} B_3 h_t$ (recall that B_3 is positive), and of ε^{D}_{t+1} it is given by $\varepsilon^{D}_{t+1} E_t \left(PD_{t+1} \right)$.

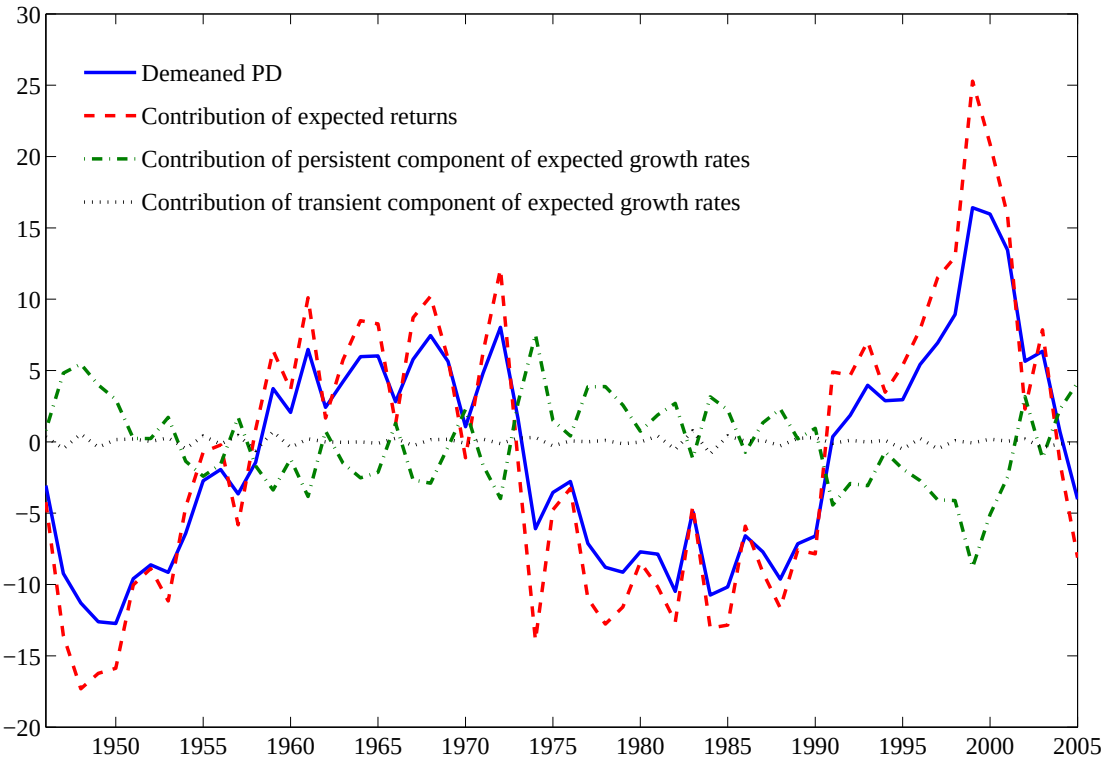


Figure 5.9: Price-dividend decomposition
We decompose the demeaned price-dividend ratio, given by $(PD_t - A)$, into the contribution of expected returns, given by $B_1\hat{\mu}_t$, the contribution of the persistent component of expected dividend growth rates $B_2\hat{g}_{1t}$, and the transient component of expected dividend growth rates given by $B_3\hat{g}_{2t}$.

Chapter 6

The Cross-section of Managerial Ability and Risk Preferences

Abstract

I use structural portfolio management models to study the joint cross-sectional distribution of managerial ability and risk preferences using manager-level data. The economic restrictions following from theory imply that (i) fund alphas reflect the managers ability and risk preferences and that (ii) information in second moments of fund returns can be used to estimate both attributes. The estimation relies on a novel framework to empirically analyze dynamic portfolio-choice models. The findings are twofold. First, the restrictions implied by recently-proposed models of managerial preferences are strongly rejected in the data. Second, introducing relative-size concerns into the managers objective delivers plausible estimates and is formally favored over the standard models and reduced-form performance regressions. Based on this model, I document large and skewed heterogeneity in risk preferences, but less dispersion in ability. Risk aversion and managerial ability are highly positively correlated. Finally, ignoring heterogeneity can lead to large welfare losses for an individual investor who allocates capital to actively-managed mutual funds.

6.1 Introduction

An investor's decision to allocate capital to actively-managed funds relies on the premise that mutual fund managers are endowed with skills that enable them to outperform a passive investment strategy. This premise has spurred a large literature aimed at measuring managerial ability and at characterizing the cross-sectional distribution of managerial talent.[1] The recent view contends that there is a small fraction of managers who are able to significantly recuperate fees and expenses.[2] However, ever since Jensen (1968), the typical approach is to rely on performance regressions to measure skill. In such performance regressions, mutual fund returns in excess of the short rate are regressed on a constant and

[1] Jensen (1968), Gruber (1996), and Carhart (1997).

[2] Kosowski, Timmermann, Wermers, and White (2006), Cremers and Petajisto (2007), Kacperczyk, Sialm, and Zheng (2005), Elton, Gruber, and Blake (2007).

a set of excess benchmarks returns. The intercept of the performance regression, the manager's alpha, is then taken as a measure of ability.[3] This reduced-form approach ignores that fund returns are the outcome of a portfolio management problem. In structural portfolio management models, the manager's ability shapes her investment opportunity set, while the manager's preferences determine which portfolio is chosen along this investment opportunity set. The fact that a manager chooses along the investment opportunity set that depends on her ability points to two important dimensions of heterogeneity: managerial ability and risk preferences. It turns out that the standard alpha reflects both managerial ability, as defined by the price of risk on the active portfolio,[4] and risk preferences. I show that the restrictions implied by structural portfolio management models can be used to disentangle both attributes. As such, this paper is the first to impose the economic restrictions following from theory to recover the joint cross-sectional distribution of managerial ability and risk preferences.

The controversy surrounding the existence of managerial ability is largely the result of inefficient inference.[5] It is well known that averaging (risk-adjusted) returns over short time spans leads to noisy estimates (Merton (1980)). Hence, the estimated cross-sectional distribution of managerial ability reflects not only true heterogeneity, but also, and perhaps predominantly, estimation error. The restrictions implied by structural portfolio management models lead to much sharper estimates of managerial ability and risk aversion because they exploit information in the volatility of fund returns and in the covariance of fund returns with benchmark returns. By combining these estimates, I recover the cross-sectional distribution of the standard performance measure, alpha. Figure 6.2 displays the cross-sectional distribution of fund alphas following from performance regressions (top panels) and structural estimation (bottom panels), both before (left panels) and after fees and expenses (right panels).

The distribution following from the structural model displays considerably less dispersion; its variance is three times smaller than in the case of standard performance regressions. For 13% of the managers, I find that their alpha is significantly higher than fees and expenses at the 5% significance level while this number drops to 9% if I would rely on standard

[3]Alternatively, risk adjustments are performed by comparing fund returns to portfolios with matched characteristics (Daniel, Grinblatt, Titman, and Wermers (1997)). Also in this case, risk-adjusted returns are averaged to gauge the manager's skill.

[4]DeTemple, Garcia, and Rindisbacher (2003) and Munk and Sørensen (2004) show that the manager's investment opportunity set can be summarized by the instantaneous short rate and prices of risk in a continuous-time economy. Nielsen and Vassalou (2004) show that if an investor has to select one fund, then she will prefer the one that has the highest price of risk. This makes the price of risk on the active portfolio a natural measure of ability motivated by portfolio-choice theory.

[5]Pástor and Stambaugh (2002b), Lynch and Wachter (2007a), and Lynch and Wachter (2007b) propose to use longer samples of benchmark returns to sharpen the estimates. Managerial ability is still estimated by averaging risk-adjusted returns over short periods.

performance regressions.

As a point of departure, I analyze two existing models of managerial preferences. In the first model, the manager derives utility from mutual fund returns in excess of a benchmark.[6] The manager's ability and her risk preferences can be recovered from the fund's beta (passive risk) and the amount of residual risk (active risk), and together imply its alpha. The resulting estimates of ability are precisely measured, but implausibly high. The implied distribution for alpha ranges from 6% to 12% on an annual basis. In the second model, the manager derives utility from assets under management, which is motivated by her compensation scheme (Elton, Gruber, and Blake (2003)). Assets under management fluctuate because of internal growth (mutual fund returns) and external growth (performance-sensitive fund flows). In addition, stellar (below-par) performance may trigger promotion (demotion) to a larger (smaller) fund.[7] It turns out that such incentives have little impact on the manager's optimal strategies in the relevant range of risk aversion. The manager therefore acts as if she cares only about internal growth. The resulting model has its own problems: the manager's risk aversion is mechanically centered around the Sharpe ratio of the benchmark return divided by its volatility. This has the undesirable consequence that the risk aversion of a given manager who controls multiple funds in different styles changes in a predictable manner across styles. As such, the resulting estimates no longer reflect risk aversion. The first important finding of this paper is therefore that the cross-equation restrictions implied by standard models of delegated management lead to economically implausible estimates of either managerial ability or risk aversion.

I propose an alternative model of managerial preferences, which imputes a concern for the relative position in the cross-sectional asset distribution into the preferences of the manager. I call this position the "fund's status." Managers are concerned about the amount of assets they have under management relative to their peers, and derive additional utility when they control larger funds. The status model nests the two standard models. I allow for different curvature parameters over assets under management and over fund status. The former preference parameter can be interpreted as risk aversion over passive risk while the latter measures risk aversion over active risk. The idea of different risk aversion parameters resonates with the views of Bob Litterman:

> *Some authors observing this perplexing use of active risk have concluded that investors actually have two separate risk aversions, one with respect to market risk, and another,*

[6] This model has been studied in Roll (1992), Becker, Ferson, Myers, and Schill (1999), Chen and Pennacchi (2007), and Binsbergen, Brandt, and Koijen (2007).

[7] Brown, Harlow, and Starks (1996), Chevalier and Ellison (1997), Sirri and Tufano (1998), Chevalier and Ellison (1999b), Busse (2001), Goriaev, Nijman, and Werker (2005), Cuoco and Kaniel (2006), Basak, Pavlova, and Shapiro (2007b), Basak, Pavlova, and Shapiro (2007a), Chen and Pennacchi (2007), Dangl, Wu, and Zechner (2007), Hu, Kale, Pagani, and Subramaniam (2007), Hugonnier and Kaniel (2007), and Chapman and Xu (2007) study the role of direct and indirect incentives.

> *much higher risk aversion, to active risk. . . . Alternatively, one might explain the active risk puzzle to be the result of an aversion on the part of fund managers to taking career (sometimes called "peer") risk that would arise if they were to create the opportunity for too much underperformance relative to their strategic benchmark. In this case, two different risk aversions make sense because they apply to very different risks faced by different individuals.*[8]

The two curvature parameters and the fund's status together imply an estimate for the manager's coefficient of relative risk aversion. Unlike the vast majority of models of delegated portfolio management, the status model is not homogenous in assets under management. It therefore predicts different risk-taking behavior for small and large funds. The two key consequences are that larger funds take on less active risk and that fund alphas are negatively related to fund size. Both are stylized facts documented in the empirical mutual fund literature, and neither can be explained by existing models.[9] This lends further credibility to the utility specification. The status model also makes contact with models where households derive utility from their position in the wealth or consumption distribution.[10] If status concerns matter at all, the mutual fund industry is likely to be the environment where such concerns are the most powerful and therefore easiest to identify.[11]

The status model leads to a plausible cross-sectional distribution of managerial ability and risk aversion as summarized in Figure 6.3. The horizontal axis displays the implied coefficient of relative risk aversion, and the vertical axis shows the price of risk on the active portfolio, which is my measure of ability. The median coefficient of relative risk aversion equals 2.51, its mean 5.16, and its standard deviation 7.69. The median price of risk on the active portfolio (to be read as a Sharpe ratio) equals .14, the mean is .28, and the standard deviation .38. Both distributions are right-skewed. In addition, managerial ability and risk aversion are highly positively correlated; their unconditional correlation is about 80%. Skilled managers are likely to be conservative. I show that this correlation is consistent with selection effects that arise when managers have a riskless outside option. Less talented managers only opt into the actively-managed mutual fund industry if they are sufficiently aggressive.

[8] *"The Active Risk Puzzle: Implications for the Asset Management Industry,"* Goldman Sachs Asset Management Perspectives, March 2004.

[9] Chevalier and Ellison (1999b) document the relation between fund size and risk-taking and Chen, Hong, Huang, and Kubik (2004) the link between fund performance and fund size.

[10] See Robson (1992), Zou (1994), Zou (1995), Bakshi and Chen (1996), Carroll (2000), Becker, Murphy, and Werning (2005), and Roussanov (2007). These models are referred to as "spirit-of-capitalism" models. It also relates to the external habit model of Campbell and Cochrane (1999) and Shore and White (2006) and to the matching models with status concerns of Cole, Mailath, and Postlewaite (2001). Goel and Thakor (2005) study the implications of status concerns for corporate investment decisions.

[11] One hypothesis is that status concerns are hard-wired in the manager's preferences. Alternatively, rank concerns may arise due to strategic interaction among fund managers (Basak and Makarov (2007)). Sirri and Tufano (1998) discuss the importance of mutual fund rankings for fund flows.

There are interesting differences in the joint distribution across the various investment styles. Given the attention asset pricing devotes to market capitalization and book-to-market ratios, it is interesting to compare the large/value and small/growth styles. The left panel of Figure 6.4 displays the estimated cross-sectional density of the coefficient of relative risk aversion for both investment styles. The median growth manager is more aggressive (median risk aversion equals 1.49) than the median value manager (median equals 3.95). The density of growth managers has much fatter tails: a substantial share of growth managers display considerably higher risk aversion than value managers. For the same groups of managers, the right panel of Figure 6.4 depicts the cross-sectional density of managerial ability. The average growth manager is more skilled, and this ordering prevails for higher-ranked managers in the tails. More generally, I analyze how cross-sectional variation in risk aversion and ability relates to observable characteristics. I find that ability is negatively related to fund size[12] and stock holdings, and positively related to the manager's tenure and asset turnover. Risk aversion is negatively related to fund size, expenses,[13] and stock holdings. In both cases, observables only account for a limited fraction of cross-sectional variation, leaving considerable unobserved heterogeneity.

The estimates of managerial ability and risk aversion follow from joint assumptions about the financial market and the manager's preferences. I formally show that the status model significantly improves upon the two standard models of delegated portfolio management. Perhaps more importantly, the status model is favored over performance regressions. This implies that the conditional distribution of the status model provides a better description of fund returns than performance regressions for which the conditional and unconditional distribution coincide. Therefore, the status model is able to capture important dynamics of mutual fund strategies that performance regressions cannot.

The average coefficient of relative risk aversion across managers varies over time due to variation in the amount of assets under management and variation in the cross-sectional asset distribution. Given the link that exists between risk aversion and the equity risk premium in equilibrium asset pricing models (Campbell and Cochrane (1999)), it is interesting to relate both time series. I measure the equity risk premium using the apparatus developed in Binsbergen and Koijen (2007). The time-series variation in risk aversion that I estimate from the universe of mutual fund managers tracks the equity risk premium; their correlation is 62% (see Figure 6.5).

[12]Chen, Hong, Huang, and Kubik (2004) document this negative relation for fund alphas, which are a non-linear function of preferences and ability. I show that even after correcting for heterogeneity in preferences, managerial ability relates negatively to fund size.

[13]The relation between expenses and both managerial ability and risk aversion is consistent with the equilibrium model of Berk and Green (2004) because the implied fund alphas relate positively to ability and negatively to risk aversion.

In conclusion, the second important finding of this paper is that introducing relative-size concerns into the manager's objective delivers plausible estimates of managerial ability and risk aversion, and is formally favored over the standard models and over reduced-form performance regressions.

I quantify the economic importance of heterogeneity by solving for the optimal allocation to the style benchmark and actively-managed mutual funds.[14] An investor who ignores heterogeneity incurs a loss in certainty-equivalent wealth of up to 4% per annum, depending on the investor's risk preferences. An investor who relies on reduced-form performance regressions to quantify heterogeneity still experiences a loss that can exceed 1% per annum. This underscores the economic importance of accounting for heterogeneity in ability and risk preferences using efficient inference methods.

I develop a novel econometric approach to bring the models to the data. The finance literature typically restricts attention to infinite-horizon models, but they are inappropriate for the problem at hand. Because the optimal policies in dynamic finite-horizon models are often unknown in closed-form, one has to rely on numerical dynamic programming. Estimating structural parameters in combination with a finite-horizon dynamic programming method is computationally (nearly) infeasible. The problem gets worse with multiple assets. The main technical contribution of this paper is to develop an estimation method that relies on the martingale method for continuous-time models in complete markets (Cox and Huang (1989)). The estimation method provides a powerful tool to formulate the likelihood and enables me for the first time to estimate dynamic finite-horizon models. One additional advantage is that the computational burden is independent of the number of assets. Koijen (2007) explains the method in a simple model and illustrates its accuracy. The method may well prove useful to estimate (i) the cross-sectional distribution of managerial ability and risk preferences in the hedge fund industry (Panageas and Westerfield (2007)), (ii) dynamic games (Basak and Makarov (2007)), and (iii) corporate finance models that can be solved by martingale methods. Since the method is likelihood-based, it can be used with both classical or Bayesian estimation procedures.[15] Finally, to estimate the models, I construct a manager-level database that covers the period 1992.1 to 2006.12.[16] Manager-fund combinations are

[14] Admati and Pfleiderer (1997), Vayanos (2003), and Binsbergen, Brandt, and Koijen (2007) consider the case in which only the risk preferences of the manager are unknown.

[15] Baks, Metrick, and Wachter (2001), Pástor and Stambaugh (2002a), and Avramov and Wermers (2006) use Bayesian methods to derive the optimal allocation to actively-managed funds. The approach developed in this paper can extend these studies in at least two directions. First, the investor forms prior beliefs over the coefficients of a performance regression, ignoring the cross-equation restrictions. The economic restrictions can discipline the set of viable priors. Second, the investor learns about ability from first moments, which is inefficient. By taking a structural approach, the investor can learn more efficiently. This increases the discrepancy between the predictive density and the investor's prior views.

[16] Other studies that construct a manager-level database are Baks (2006), Evans (2007), and Kacperczyk and Seru (2007).

allocated to one of nine investment styles that differ by size and book-to-market orientation.

While the finance literature has been focussing primarily on quantifying managerial ability, recovering risk preferences is an important question in the economics literature. Most estimates follow from game shows (Gertner (1993), Metrick (1995), and Assem, Baltussen, Post, and Thaler (2007)), horse races (Jullien and Salanié (2000)), car insurance markets (Cohen and Einav (2007)), labor supply decisions (Chetty (2006)), hypothetical income gambles (Kimball, Sahm, and Shapiro (2007)), or experiments.[17] Most closely related is the consumption-based asset pricing literature, which uses the household's Euler condition to estimate preference parameters. I propose to use the first-order conditions of fund managers to estimate ability and risk preferences. The mutual fund industry provides a great laboratory wherein to recover risk preferences for at least two important reasons.[18] First, fund managers routinely take decisions under uncertainty. This implies that I observe a series of decisions by the same manager to estimate risk preferences. Second, the decisions made by the manager involve non-trivial sums of money and have non-trivial implications for the manager's career. In addition, as argued by Cohen and Einav (2007), it is important to estimate risk preferences in the environment in which they will be applied. Estimates of risk aversion are of considerable practical relevance in the context of an investor's decision to allocate money to actively-managed funds and in the context of mutual fund valuation.[19]

The paper proceeds as follows. Section 6.2 describes the data. I provide details on the financial market model in Section 6.3. Section 6.4 introduces two standard models of delegated portfolio management and derives the cross-equation restrictions that are implied by theory. Next, Section 6.5 discusses the econometric approach and Section 6.6 provides the empirical results for the benchmark models. Section 6.7 introduces the status model and Section 6.8 contains the empirical results for this model. I study the economic importance of heterogeneity for asset allocation decisions in Section 6.9. Finally, Section 6.10 concludes.

6.2 Data

Data sources I combine data from three sources. First, monthly mutual-fund returns come from the Center for Research in Securities Prices (CRSP) Survivor Bias Free Mutual

[17] Andersen, Harrison, Lau, and Rutstrom (2005) discuss the applicability of results obtained from experiments to real-life settings.

[18] Becker, Ferson, Myers, and Schill (1999) estimate a structural model of delegated management, but they impose the restrictions only on the benchmark allocation.

[19] Boudoukh, Richardson, Stanton, and Whitelaw (2004), Huberman (2007), and Dangl, Wu, and Zechner (2007) develop models of mutual fund valuation. These models can be extended with the preference specifications in this paper to study how heterogeneity in managerial ability and risk preferences impact fund valuation.

Fund Database. The CRSP database is organized by fund rather than by manager, but contains manager's names starting in 1992. I use the identity of the manager to construct a manager-level database. The sample consists of monthly data over the period from January 1992 to December 2006. Data on the Fama and French size (SMB) and book-to-market (HML) portfolios, Carhart (1997)'s momentum factor, and the short-rate data also come from CRSP. Second, manager-fund combinations are allocated to investment styles. I consider different approaches for robustness (discussed below). In one approach, I use the benchmark mapping from Cremers and Petajisto (2007).[20] Third, benchmark returns are obtained from Datastream.

Sample selection I apply several screens to the mutual-fund data to obtain a sample of active domestic-equity portfolio managers. I first classify the funds by the investment objectives "Small company growth," "Other aggressive growth," "Growth," "Growth and income," and "Maximum capital gains" using the Wiesenberger, ICDI, or Strategic Insight Codes (Pástor and Stambaugh (2002b)). All funds that cannot be classified are omitted from the sample.[21] I drop funds that have an average total equity position (common plus preferred stock) smaller than 80% in order to focus on all-equity funds. I also drop fund years for which the total net assets are smaller than $10 Million. I omit observations for which the manager's name is missing and the years for which no information on returns or total net assets is available. I only include fund years for which the fund is "Active" in the nomenclature of Cremers and Petajisto (2007).

Several screens are specific to the manager-level database. First, manager names in the CRSP database can take three forms: a manager/management team is (i) fully identified, (ii) partly identified, or (iii) fully anonymous. For the partly identified or anonymous management teams, I consider separately each team that manages a different fund.[22] This presumably overstates the number of anonymous management teams in the mutual fund industry, but there is no alternative way to match such (partly) unidentified teams. I focus for most of the analysis on funds for which the manager/management team is fully identified. Managers are matched on the basis of their names. Names in the CRSP database are, however, often misspelled and abbreviated in different ways. I first use a computer algorithm that detects commonly made errors. I then manually check all funds carefully and code them consistently. Further, the manager's starting date in the CRSP database is

[20] I am grateful to Martijn Cremers and Antti Petajisto for sharing the data on the benchmark mapping. If ω_{it} denotes the weight at time t in stock i, and ω_{it}^b the corresponding weight in benchmark b, then the active share relative to benchmark b is defined as: $AS_t(b) \equiv \frac{1}{2}\sum_{i=1}^{N} \left|\omega_{it} - \omega_{it}^b\right|$. In addition, they define a fund to be active when the active share exceeds 30% and when the name does not contain "Index" or "Idx."

[21] This selection excludes international funds, bond funds, money market funds, sector funds, and funds that do not hold the majority of their securities in US equity.

[22] Massa, Reuter, and Zitzewitz (2007) study the role of anonymous teams in delegated asset management.

subject to substantial measurement error (Baks (2006)). I remove a fund from a manager's career profile when the starting date contains inconsistencies (Baks (2006) and Kacperczyk and Seru (2007)).

US mutual funds typically have multiple share classes associated with different fee structures.[23] Consistent with the literature, I merge different share classes: I construct value-weighted returns, loads, expense ratios, and 12B-1 fees,[24] fraction in stocks, and cash using the total net assets of the different share classes to construct the weights. I select the other variables from the share class that has the highest total net assets (Cremers and Petajisto (2007)). Finally, several managers manage multiple funds at the same time. For funds that are compared to the same benchmark, I merge the return using the fund's total net assets to weigh them. When funds operate in different styles, I keep them as separate observations.

Mutual fund costs CRSP mutual fund returns are net fees and expenses, but computed before back- and front-end loads. To focus on true managerial ability, I compute gross returns by adding back expense ratios in line with Wermers, Yao, and Zhao (2007). I sum the annual expense ratio divided by 12 to each monthly return in a particular year.

Benchmark selection The prime motivation for benchmarking is to disentangle managerial skill and effort from the reward of following passive strategies.[25] Benchmark selection is notoriously difficult, regardless of whether one relies on regression techniques, matched characteristics, or self-reported benchmarks.[26] I employ two procedures to identify the benchmark for each manager-fund combination; one is regression-based and the other is holdings-based. In the first approach, I regress mutual fund returns on benchmark returns, both in excess of the short rate, and select the benchmark that maximizes the R-squared. Alternatively, I use the method of Cremers and Petajisto (2007), which selects the benchmark that minimizes the active share of the fund. This approach leads to a benchmark that has the highest overlap with a fund's holdings. In this paper, I report all results for the regression-based approach. The main results are insensitive to the benchmark selection methodology.

I consider a set of nine benchmarks that are distinguished by their size and value orientation. For large-cap stocks, I use the S&P 500, Russell 1000 Value, and Russell 1000 Growth; for mid-cap stocks, I take the Russell Midcap, Russell Midcap Value, and Russell Midcap

[23] Nanda, Wang, and Zheng (2007) study the share-class structure of mutual funds.

[24] 12B-1 fees cover expenses related to selling and marketing shares, see Barber, Odean, and Zheng (2005).

[25] Admati and Pfleiderer (1997), Binsbergen, Brandt, and Koijen (2007), and Basak, Pavlova, and Shapiro (2007a) discuss advantages and disadvantages of benchmarking.

[26] Chan, Dimmock, and Lakonishok (2006) and Sensoy (2007) provide results on different benchmark selection methodologies. Brown and Goetzmann (1997) provide an interesting alternative regression-based selection methodology.

Growth; for small-cap stocks, I select the Russell 2000, Russell Value, and Russell 2000 Growth. The style indexes are taken from Russell, in line with Chan, Chen, and Lakonishok (2002) and Chan, Dimmock, and Lakonishok (2006).[27]

Summary statistics The sample consists of 3,694 unique manager-fund combinations of 3,163 different managers who manage 1,932 different mutual funds. For 1,273 manager-fund combinations I have more than three years of data available. I impose a minimum data requirement of three years to estimate all models so that performance regressions deliver reasonably accurate estimates. The left panel of Table 6.3 displays the allocation of manager-fund combinations to the nine styles for the full sample, the right panel for manager-fund combinations for which at least 3 years of data is available. One fifth of the managers are compared to the S&P 500.[28] The benchmarks are relatively equally divided across the size dimension, with a small tilt towards large-cap benchmarks. The majority of the large-cap funds are neutral in the value-dimension, but the medium- and small-cap funds are predominantly growth-oriented.

Table 6.4 provides summary statistics for the total net assets under management (TNA), total net assets of the fund family (as defined in Chen, Hong, Huang, and Kubik (2004)), family size (the number of funds that belong to the fund family), expense ratio, 12B-1 fees, the total load (the sum of maximum front-end load fees and maximum deferred and rear-end load fees), cash holdings as reported by the fund, stock holdings as reported by the fund (the sum of common and preferred stock), manager's tenure, fund age, and annual turnover. The summary statistics are broadly consistent with prior studies.

6.3 Financial market

The manager's asset menu contains three assets. The first asset is a cash account that trades at price S_t^0. The cash account earns a constant interest rate r and its dynamics satisfy:

$$\mathrm{d}S_t^0 = S_t^0 r \mathrm{d}t. \tag{6.1}$$

[27]The correlation with the corresponding style index from Standard and Poor's is in all cases higher than 96.5%, measured over the full sample period.

[28]Elton, Gruber, and Blake (2003) find that managers that have explicit incentives in their compensation schemes are compared to the S&P 500 in 44% of the cases. This suggests that this benchmark is either more popular with managers who receive incentive compensation, or that managers deviate from their stated objectives. Sensoy (2007) provides evidence that managers deviate from the benchmarks reported in the prospecti of these funds.

The second asset is the benchmark portfolio with price S_t^B:

$$\mathrm{d}S_t^B = S_t^B \left(r + \sigma_B \lambda_B\right) \mathrm{d}t + S_t^B \sigma_B \mathrm{d}Z_t^B, \tag{6.2}$$

where λ_B is the price of risk, σ_B the standard deviation of the benchmark portfolio, and Z_t^B a standard Brownian motion. The coefficients are assumed to be constant during the investment period.[29] Third, manager i can trade a manager-specific active portfolio with price S_{it}^A. Without loss of generality, I assume that the active asset does not carry any systematic risk. The dynamics of the active portfolio read:

$$\mathrm{d}S_{it}^A = S_{it}^A \left(r + \sigma_{Ai} \lambda_{Ai}\right) \mathrm{d}t + S_{it}^A \sigma_{Ai} \mathrm{d}Z_{it}^A. \tag{6.3}$$

I take the price of risk on the active portfolio, λ_{Ai}, as the measure of managerial ability.[30] σ_{Ai} denotes the volatility of the active portfolio.

It would be straightforward to extend the model with a set of passive portfolios that can easily be replicated by managers, like momentum, which are typically not considered to reflect skill.

Benchmark portfolio, assets under management, and state-price density The benchmark portfolio is given by the two-dimensional vector $V = (v, 0)'$ of portfolio weights. The remainder, $1 - v$, is allocated to the cash account.[31] The value of the benchmark at time t is denoted by B_t. The benchmark dynamics read:

$$\mathrm{d}B_t = B_t \left(r + v\sigma_B \lambda_B\right) \mathrm{d}t + B_t v \sigma_B \mathrm{d}Z_t^B. \tag{6.4}$$

Assets under management at time t, A_{it}, evolve according to:

$$\begin{aligned} \mathrm{d}A_{it} &= A_{it} \left(r + x_{it}^B \sigma_B \lambda_B + x_{it}^A \sigma_{Ai} \lambda_{Ai}\right) \mathrm{d}t + A_{it} x_{it}^B \sigma_B \mathrm{d}Z_t^B + A_{it} x_{it}^A \sigma_{Ai} \mathrm{d}Z_{it}^A \\ &= A_{it} \left(r + x_{it}' \Sigma_i \Lambda_i\right) \mathrm{d}t + A_{it} x_{it}' \Sigma_i \mathrm{d}Z_{it}, \end{aligned} \tag{6.5}$$

[29] Koijen (2007) discusses extensions of the econometric framework to accommodate time-varying interest rates and prices of risk during the investment period. For tractability, I assume the parameters to be constant during the investment period of one year. I update the short rate on an annual basis.

[30] The model can easily be set up by allowing the manager to trade J stocks with different prices of risk. However, in all models I consider, the manager will perfectly diversify the active portfolio leading to a single active portfolio (see Chen and Pennacchi (2007) and Basak, Pavlova, and Shapiro (2007b)). The formation of the active portfolio becomes important in models of costly information acquisition (Van Nieuwerburgh and Veldkamp (2007)). In this case, the active portfolio will not be perfectly diversified, as costly learning capacity is allocated to a few stocks only.

[31] I assume fixed benchmark weights. Binsbergen, Brandt, and Koijen (2007) show that benchmarks with constant weights can alleviate most efficiency losses that arise in a decentralized investment management environment. Basak, Pavlova, and Shapiro (2007a) provide a similar result for a manager that shifts risk in response to incentives. Both studies suggest there is little need for dynamic benchmarks.

where x_{it}^B and x_{it}^A are the fractions invested in the benchmark and active portfolio, $x_{it} \equiv \left(x_{it}^B, x_{it}^A\right)'$, $\Sigma_i \equiv \text{diag}(\sigma_B, \sigma_{Ai})'$, $\Lambda_i \equiv (\lambda_B, \lambda_{Ai})'$, and $Z_{it} \equiv \left(Z_t^B, Z_{it}^A\right)'$. The asset dynamics excludes fund flows from outside investors, which I will discuss in detail in Section 6.4.2.

The state-price density at time t of manager i is denoted by φ_{it}. The state-price density plays a key role in the econometric approach and its dynamics satisfy:

$$\mathrm{d}\varphi_{it} = -\varphi_{it} r \mathrm{d}t - \varphi_{it} \Lambda_i' \mathrm{d}Z_{it},\ \varphi_{0i} = 1. \tag{6.6}$$

I will omit the subscripts i for the remainder of the paper for notational convenience. However, note that λ_A, σ_A, and Z_t^A, and correspondingly S_t^A, x_t, and φ_t, are manager-specific. The remaining parameters are common across all managers in a particular style.

6.4 Standard models of delegated management

As a point of departure, I consider two standard models of delegated portfolio management that have been suggested in the literature in Section 6.4.1 and 6.4.2. Section 6.4.3 derives the cross-equation restrictions that are implied by theory.

6.4.1 Relative-return preferences

The first model assumes that the manager derives utility from assets under management relative to the value of the benchmark:

$$\max_{(x_s)_{s\in[0,T]}} E_0 \left[\frac{1}{1-\gamma} \left(\frac{A_T}{B_T} \right)^{1-\gamma} \right], \tag{6.7}$$

where γ is the coefficient of relative risk aversion. This model captures, in a reduced-form, that the manager's performance and ultimately her compensation is relative to a benchmark.[32] The optimal strategy is a constant-proportions strategy (Binsbergen, Brandt, and Koijen (2007)):

$$x^\star = \frac{1}{\gamma} \left(\Sigma\Sigma'\right)^{-1} \Sigma\Lambda + \left(1 - \frac{1}{\gamma}\right) V. \tag{6.8}$$

It combines the mean-variance portfolio and the benchmark portfolio. The two portfolios are weighted by the coefficient of relative risk aversion. Consistent with the standard interpretation in the investment industry, infinitely risk-averse agents (that is, $\gamma \to \infty$) hold the

[32] Binsbergen, Brandt, and Koijen (2007) and Chen and Pennacchi (2007) provide some further motivation for these preferences.

benchmark ($x^\star = V$).

6.4.2 Preferences for assets under management

The standard model The second model assumes that the manager derives utility from assets under management:

$$\max_{(x_s)_{s\in[0,T]}} E_0\left[\frac{1}{1-\gamma}A_T^{1-\gamma}\right], \tag{6.9}$$

where γ denotes the coefficient of relative risk aversion. These preferences are motivated by the observation that most managers are compensated by a fraction of the assets under management (Deli (2002)). The optimal strategy reads:

$$x^\star = \frac{1}{\gamma}\left(\Sigma\Sigma'\right)^{-1}\Sigma\Lambda, \tag{6.10}$$

which is also of the constant-proportions type.

Preferences, career concerns, and fund flows Assets under management fluctuate due to investment returns (internal growth), but also due to fund flows and promotion or demotion of the fund manager (external growth). Both performance-sensitive fund flows and career concerns may motivate the manager to deviate from the optimal strategy in (6.10). It is well known from empirical studies that new capital flows disproportionally to funds with stellar performance,[33] which results in an increasing and convex flow-performance relationship. In addition, exceptional (below-par) performance can lead to promotion (demotion) to a larger (smaller) fund. I analyze the importance of these incentives using the calibration of Chapman and Xu (2007). They calibrate promotion/demotion probabilities to observed career events and estimate the flow-performance relationship.[34] Appendix 6.B uses this model to study the interaction between incentives and risk aversion. I show that in the relevant range of risk aversion, managerial incentives are not powerful enough to distort the optimal strategy. I therefore abstract from such incentives in the main text.

[33]See for instance Brown, Harlow, and Starks (1996), Chevalier and Ellison (1997), and Sirri and Tufano (1998). Lynch and Musto (2003), Berk and Green (2004), and Hugonnier and Kaniel (2007) develop theoretical models to rationalize the relation between performance and fund flows.

[34]Hu, Hall, and Harvey (2000) and Baks (2006) also estimate promotion and demotion probabilities. I use the calibration of Chapman and Xu (2007) because their calibration covers most closely the sample period I study (1994 to 2006).

6.4.3 Cross-equation restrictions implied by structural models

The bulk of the performance literature averages risk-adjusted returns to obtain an estimate of managerial ability. This means that a few years of data are used to estimate a mean return, a notoriously noisy approach.[35] The key difference in this paper is to use the optimality conditions of the manager's portfolio problem to uncover managerial ability. In consumption-based asset pricing, it is common to use the household's Euler condition to estimate preference parameters. The cross-equation restrictions implied by the standard models show that both ability and risk aversion can be estimated from second moments of fund returns. I use the first model to illustrate the restrictions.

I start from a standard performance regression, formulated in continuous time:

$$\frac{\mathrm{d}A_t}{A_t} - r\mathrm{d}t = \alpha\mathrm{d}t + \beta\left(\frac{\mathrm{d}S_t^B}{S_t^B} - r\mathrm{d}t\right) + \sigma_\varepsilon \mathrm{d}Z_t^A, \tag{6.11}$$

which, using (6.2), is equivalent to:

$$\frac{\mathrm{d}A_t}{A_t} = (r + \alpha + \beta\sigma_B\lambda_B)\,\mathrm{d}t + \beta\sigma_B\mathrm{d}Z_t^B + \sigma_\varepsilon\mathrm{d}Z_t^A. \tag{6.12}$$

The parameters α, β, and σ_ε are manager-specific; λ_B and σ_B are common to all managers.

The relative-return preferences in (6.7) lead to the optimal portfolio in (6.8), which I substitute into (6.5) to obtain the optimal asset dynamics:

$$\frac{\mathrm{d}A_t}{A_t} = \left(r + \frac{\lambda_B^2}{\gamma} + \left(1 - \frac{1}{\gamma}\right)v\sigma_B\lambda_B + \frac{\lambda_A^2}{\gamma}\right)\mathrm{d}t + \left(\frac{\lambda_B}{\gamma} + \left(1 - \frac{1}{\gamma}\right)v\sigma_B\right)\mathrm{d}Z_t^B + \frac{\lambda_A}{\gamma}\mathrm{d}Z_t^A, \tag{6.13}$$

where λ_A and γ are manager-specific, and v is common to all managers. The cross-equation restrictions implied by the structural model follow from matching the drift and diffusion terms in (6.12) and (6.13):

$$\alpha = \lambda_A^2/\gamma, \tag{6.14}$$

$$\beta = \frac{\lambda_B}{\gamma\sigma_B} + \left(1 - \frac{1}{\gamma}\right)v, \tag{6.15}$$

$$\sigma_\varepsilon = \lambda_A/\gamma. \tag{6.16}$$

[35] This noise has motivated researchers to form portfolios based on observable characteristics to identify quality managers. These include the portfolios' active share (Cremers and Petajisto (2007)), similarities in portfolio holdings (Cohen, Coval, and Pástor (2005)), measures of concentration in portfolio holdings (Kacperczyk, Sialm, and Zheng (2005)), or their reliance on public information (Kacperczyk and Seru (2007)). By pooling managers cross-sectionally, the precision of the estimates increases.

The right-hand side of (6.15) and (6.16) identifies the two manager-specific structural parameters, λ_A and γ. They are identified off the fund's beta:[36]

$$\beta = \frac{\text{Cov}\left(\frac{\mathrm{d}A_t}{A_t}, \frac{\mathrm{d}S_t^B}{S_t^B}\right)}{\text{Var}\left(\frac{\mathrm{d}S_t^B}{S_t^B}\right)}, \tag{6.17}$$

and residual risk:

$$\sigma_\varepsilon^2 = \text{Var}\left(\frac{\mathrm{d}A_t}{A_t}\right) - \beta^2 \text{Var}\left(\frac{\mathrm{d}S_t^B}{S_t^B}\right), \tag{6.18}$$

Equation (6.14) can be used to restrict α:[37]

$$\alpha = \sigma_\varepsilon^2 \left(\frac{\lambda_B/\sigma_B - v}{\beta - v}\right). \tag{6.19}$$

Recall that β and σ_ε are manager-specific, whereas λ_B, σ_B, and v are common to all managers. This results in an estimate of the manager's alpha via (6.19) that relies solely on information in second moments.

The typical approach in the literature is to estimate α, β, and σ_ε separately. The resulting estimate for α is based on information in average (risk-adjusted) fund returns. As it turns out, this is the most inefficient moment to use. The likelihood-based estimation procedure in Section 6.5 efficiently combines information from the average and the volatility of fund returns as well as the covariance of fund returns with benchmark returns. Since likelihoods are typically much steeper in parameters that govern second moments, ability is effectively estimated using that information.

Simulation exercise The structural model implies that alpha is a performance measure that mixes information on ability (λ_A) and preferences (γ). In addition, it shows that second moments of mutual fund returns contain useful information on preferences and ability. To illustrate the benefits of imposing this cross-equation restriction, Table 6.5 provides the results of a simple simulation experiment. I simulate 2,500 sets of three years of monthly data from the model. The price of active risk takes values $\lambda_A \in \{.1, .2, .3\}$ and the coefficient of relative risk aversion takes values $\gamma \in \{2, 5, 10\}$. The market parameters correspond to the S&P 500 as the style benchmark. Panel A of Table 6.5 provides the results for the maximum-likelihood estimators of λ_A and γ. The resulting estimates are unbiased and sharp. For

[36] Formally, the covariance needs to be interpreted as the quadratic covariation, the variance as the quadratic variation, and $\mathrm{d}t = 1$.

[37] A similar restriction arises in the model of Section 6.4.2, $\alpha = \sigma_\varepsilon^2 \lambda_B/(\sigma_B \beta)$, which coincides with (6.19) if $v = 0$, that is, in case of a cash benchmark.

$\lambda_A = .2$ and $\gamma = 5$, an 80%-confidence interval for λ_A ranges from .16 to .24; for γ from 4.5 to 5.6. Panel B of Table 6.5 illustrates the efficiency gains for fund alphas. I compare the model-implied alpha (α_{ML}) to the one that follows from a performance regression (α_{OLS}). When $\lambda_A = .2$ and $\gamma = 5$, the true $\alpha = .8\%$. The 80%-confidence interval for $\alpha_{ML} = [.54\%, 1.05\%]$, whereas for $\alpha_{OLS} = [-2.15\%, 3.93\%]$. In this example, the standard deviation of α_{ML} is .78%, whereas the standard deviation of α_{OLS} is *three times* larger at 2.37%. This implies that standard performance regressions require nine times more data if the cross-equation restrictions are not imposed to deliver the same accuracy in this model. It resonates with the empirical results in Figure 6.2 in the introduction, which compares the implied estimates of alpha following from the model in Section 6.7 to the estimates of performance regressions. This illustrates that imposing the restrictions implied by theory significantly sharpens the implied estimates of α.

6.5 Econometric approach

In this section, I develop a general method to estimate the ability and preference parameters of dynamic models of delegated portfolio management in a complete-markets setting. The method is likelihood-based and can therefore be combined with both classical and Bayesian estimation procedures. Appendix 6.E contains further details and Koijen (2007) discusses a simple example to illustrate the method and its accuracy. For the models in Section 6.4, it is possible to construct estimators that are easier to implement (Appendix 6.E.1). However, because the estimates of ability and risk aversion following from both standard models are economically implausible, I generalize the preferences in Section 6.7. This model can no longer be estimated using standard techniques and requires the novel approach given in this section.

The inference problem Consider a manager who can trade the style benchmark, the active portfolio, and cash. I estimate the model using information on benchmark returns, r^{BT}, $r_t^B \equiv \log S_t^B - \log S_{t-h}^B$, and mutual funds returns, $r_t^A \equiv \log A_t - \log A_{t-h}$, with $y^T \equiv \{y_h, \ldots, y_T\}$. I take $h = 1/12$ since the model is estimated using monthly data. I set the short rate, r, equal to the average 30-day T-bill rate during the investment period, which I take to be one year. The model parameters can be grouped into financial market parameters that apply to all managers, $\Theta_B \equiv \{\lambda_B, \sigma_B\}$, and parameters that are manager-specific, $\Theta_A \equiv \{\lambda_A, \gamma\}$.[38]

I adopt a two-step procedure to estimate the model. First, I estimate the financial market

[38]The volatility of the active portfolio cannot be identified from returns data only. This parameter is, however, unimportant because it does not enter the likelihood once evaluated at the optimal strategy.

parameters that are common to all managers, Θ_B. Because asset prices follow geometric Brownian motions conditional on the short rate, the log-likelihood of r^{BT}, $\mathcal{L}(r^{BT};\Theta_B)$, is trivial to construct. In the second step, I estimate the manager-specific parameters, Θ_A, using the log-likelihood of fund returns conditional on the benchmark returns and the first-stage estimates, $\mathcal{L}(r^{AT} \mid r^{BT};\Theta_A,\hat{\Theta}_B)$. The main complication is to compute the second-stage likelihood.

While a single-step estimation would enhance the efficiency of the estimates, it would require modeling the cross-sectional correlation of active portfolio returns. The two-step procedure accommodates any cross-sectional dependence in active returns. It therefore requires less restrictive statistical assumptions, is not subject to misspecification of the correlation structure, and still results in consistent estimates. In addition, the two-step procedure saves substantially on computational time.

The conditional log-likelihood of mutual fund returns To appreciate why it is non-trivial to construct $\mathcal{L}(r^{AT} \mid r^{BT};\Theta_A,\hat{\Theta}_B)$, consider the dynamics of assets under management:

$$\mathrm{d}A_t = A_t\left(r + x_t^\star(A_t)'\Sigma\Lambda\right)\mathrm{d}t + A_t x_t^\star(A_t)'\Sigma\mathrm{d}Z_t, \tag{6.20}$$

where $x_t^\star(A_t)$ is the optimal investment strategy of the manager, which may depend on time and assets under management. There are two complications, which are related. First, the diffusion coefficient, $x_t^\star(A_t)'\Sigma$, may be time varying if the manager implements a dynamic strategy. This is the case for the model I study in Section 6.7. This often implies that the exact discretization is unknown, which leads to a discretization bias. The typical approach in the literature is to stabilize the diffusion coefficient to mitigate the discretization bias. The likelihood is then constructed via simulations (Brandt and Santa-Clara (2002) and Durham and Gallant (2002)) or series expansions of the transition density (Ait-Sahalia (2002), Ait-Sahalia (2007), and Bakshi and Ju (2005)). Second, the optimal strategy, and therefore the diffusion coefficient, is in most cases not known analytically. This implies that standard stabilization methods cannot be implemented.

One solution would be to solve the dynamic problem numerically either in discrete or continuous time.[39] This approach has, at least, two drawbacks. First, solving the discrete-time problem is computationally expensive. This stems from the fact that these dynamic models typically feature one endogenous state variable, in this case assets under management. This implies that the optimal policy needs to be constructed on a grid for each period. Second, and related, the computational costs increase exponentially in the number of assets.

[39] See Balduzzi and Lynch (1999), Brandt, Goyal, Santa-Clara, and Stroud (2005), and Koijen, Nijman, and Werker (2007a) for recent advances to solve such problems.

I can side-step these issues in computing the likelihood.

The manager's problem and the martingale method The econometric approach relies on the martingale method of Cox and Huang (1989). I first solve for the optimal terminal asset level, $A_T^\star$:

$$\max_{A_T \geq 0} E_0 \left[u \left(A_T \right) \right], \tag{6.21}$$

$$\text{s.t.} \quad E_0 \left[\varphi_T A_T \right] \leq A_0, \tag{6.22}$$

where (6.22) is the static representation of the dynamic budget constraint in (6.5). If the utility index is strictly concave, it holds that $A_T^\star = I \left(\xi \varphi_T \right)$, where ξ is the Lagrange multiplier corresponding to the budget constraint and $I(\cdot) \equiv \left(u' \right)^{-1} (\cdot)$. By no-arbitrage, time-$t$ assets under management satisfy:

$$A_t^\star(\varphi_t) = E_t \left[I \left(\xi \varphi_T \right) \frac{\varphi_T}{\varphi_t} \right], \tag{6.23}$$

which is a function of φ_t only because $(\varphi_t)_{t \geq 0}$ is Markovian. Since the utility index is strictly concave, $A_t^\star(\varphi_t)$ is invertible (Koijen (2007)). This implies that observing assets under management, or fund returns, is equivalent to observing the time series of the state-price density, φ^T. I then apply the Jacobian formula:

$$\ell \left(r_t^A \mid r_t^B, \varphi_{t-h}; \Theta_A, \Theta_B \right) = \ell \left(\varphi_t \mid r_t^B, \varphi_{t-h}; \Theta_A, \Theta_B \right) + \log \left| \left(\frac{\partial \left(\log A_t^\star - \log A_{t-h}^\star \right)}{\partial \varphi_t} \right)^{-1} \right| \tag{6.24}$$

and note that φ_{t-h} (or, equivalently, $A_{t-h}^\star$) contains all time-$(t-h)$ information needed due to the Markov property. Both terms in (6.24) are straightforward to compute. Because φ_t is log-normally distributed given φ_{t-h} and r_t^B, this involves one-dimensional Gaussian quadrature. Koijen (2007) demonstrates its accuracy for a low number of quadrature points.

In Section 6.7, I develop a model in which the utility index is not globally concave. This implies that the martingale approach cannot be applied directly. The solution proposed in the literature is to replace the original utility index with the smallest concave function that dominates it (Carpenter (2000), Cuoco and Kaniel (2006), and Basak, Pavlova, and Shapiro (2007b)), and then use standard techniques. Appendix 6.E.2 applies this approach to the model of Section 6.7. Further, I construct the standard errors using the outer-product gradient estimator. Appendix 6.F shows how to test hypotheses in dynamic models of delegated portfolio management. I use these tests to compare different nested and non-nested models.

This method results in the exact likelihood of fund returns, up to the computation of an expectation of a univariate random variable and its numerical derivative. The method is insensitive to endogenous state variables and the computational effort is independent of the number of assets in the manager's menu. The only restriction on the method is that the market is dynamically complete.[40]

6.6 Empirical results for the benchmark models

Relative-return preferences Table 6.6 displays results for the model in Section 6.4.1. The benchmark weights are set to $V = (1, 0)'$ and T is set to one year.[41] The first two columns provide summary statistics for the estimates of ability and relative risk aversion for the nine investment styles. Columns three to five show the implied coefficients of a performance regression using Equation (6.14), and columns six to eight contain the results of standard performance regressions in a continuous-time framework (Appendix 6.A).

The average coefficient of relative risk aversion is high and its distribution is right-skewed. The intuition is that mutual funds have a β that is close to one. Since λ_B/σ_B substantially exceeds one for all styles, Equation (6.15) implies that γ needs to be high. However, to match the amount of active risk that managers take, σ_ε, the price of risk needs to be high to offset the high risk aversion estimate (see Equation (6.16)). The average price of risk ranges from .64 (small/growth) to 1.75 (midcap/value). Consequently, the implied alpha is implausibly high, a result of Equation (6.14). The average estimates for alpha are between 6.14% and 11.88% per annum. The average alpha is substantially higher than the alpha following from standard performance regressions for all investment styles. As such, this model is unable to simultaneously match the fund's active and passive risk-taking and a low average risk-adjusted return.

Preferences for assets under management Table 6.7 displays the results for the model in Section 6.4.2 and has the same structure as Table 6.6. This model almost perfectly replicates the distribution of active (σ_ε) and passive (β) risk-taking. The estimates for γ and λ_A are considerably lower than for the model in Section 6.4.1. The estimates of alpha are correspondingly lower, and range from 86 basis points (bp) to 294bp. Despite the more reasonable estimates for managerial ability, the average coefficient of relative risk aversion

[40] It is theoretically possible to apply martingale techniques even in incomplete markets. See for instance He and Pearson (1991), Cvitanic and Karatzas (1992), and the application in Sangvinatsos and Wachter (2005).

[41] As an alternative, I also consider $\beta = (1 - \text{cash}, 0)$, with "cash" the average cash position in a particular year, and $\beta = (1 - \text{stock}, 0)$, with "stock" the fraction invested in common and preferred equity. The main conclusions are comparable for these alternative benchmark strategies.

tracks λ_B/σ_B and displays little dispersion. The reason is that mutual funds have, on average, a beta of one with respect to the style benchmark. To generate a unit beta, γ equals λ_B/σ_B because the fund's beta (x^B) is in this model given by $\lambda_B/(\gamma\sigma_B)$, see (6.8). This means that, by default, a value manager has a higher coefficient of risk aversion than a growth manager, on average, because the price of risk is higher and the volatility is lower for value stocks. Therefore, the estimated γ does not reflect risk aversion. To make this point more clearly, I consider a sample of managers who manage multiple funds at the same point in time (not reported). Such managers should display stable risk preferences across styles. However, it turns out that the risk aversion estimates contain a "fixed effect," captured by λ_B/σ_B. Finally, note also that the distribution is virtually symmetrical and displays very little dispersion (Table 6.7). This is at odds with Cohen and Einav (2007) and Kimball, Sahm, and Shapiro (2007), who find strong evidence in favor of right-skewed distributions.

In summary, this model generates estimates of risk aversion that are mechanically tied to that of the representative agent, leading to low dispersion in preference parameters and a "fixed effect" per asset class. That is, the average risk aversion moves in lock-step with λ_B/σ_B, which contaminates its interpretation as a coefficient of relative risk aversion.

Robustness I consider various extensions to ensure the robustness of these results. I allow for (i) time variation in risk premia that is governed by the short rate and the dividend yield (Ang and Bekaert (2007)),[42] (ii) other passive portfolios such as momentum, (iii) cash positions in the benchmark, (iv) stochastic volatility, and (v) learning about managerial ability.[43] These modifications do not alter the conclusions qualitatively. In conclusion, neither of the two standard models produces sound estimates of the joint distribution of managerial ability and risk preferences.

[42]Becker, Ferson, Myers, and Schill (1999) discuss the importance of conditioning information in market timing models.

[43]I extend the model in Section 6.4.1 to allow for the possibility that the manager does not know her ability as in Berk and Green (2004) and Dangl, Wu, and Zechner (2007). Instead, the manager starts off with a (Gaussian) prior on the price of risk and updates her views based on realized performance (Cvitanic, Lazrak, Martellini, and Zapatero (2006)). The estimation error that is taken into account increases the effective risk aversion. This implies that the manager's prior mean needs to be even higher than the estimates in absence of parameter uncertainty to reconcile active risk taking. Consequently, the estimates for the prior mean are economically implausible or the prior is very tightly centered around the maximum likelihood estimates without learning, which shuts down the learning channel. Alternatively, I use the cross-sectional distribution of the mutual fund performance to form the prior instead of estimating the prior distribution for each manager separately. A formal specification test indicates that both learning models are strongly rejected in favor of the status model in Section 6.7. Technical details and further empirical results are available upon request.

6.7 Status model for delegated portfolio management

In this section, I develop and study the main implications of a new model of delegated investment management that features status concerns on the part of the manager. Section 6.8 presents the main empirical results.

Motivation Standard models of delegated portfolio managers postulate that the manager cares only about assets under management or about performance relative to a benchmark. However, a large literature in sociology and economics[44] argues that status considerations may be important for economic behavior and financial decision-making. Given the numerous rankings of mutual funds and fund managers and their importance for fund flows (Sirri and Tufano (1998)), the mutual fund industry provides an economic environment where status concerns are clearly important. I generalize the manager's preferences so that she derives utility from both assets under management and the position of the fund in the cross-sectional asset distribution. I call the latter the fund's status. There are at least two ways to motivate the status-seeking behavior of fund managers. One hypothesis is that status concerns are hard-wired into the manager's preferences as a result of evolutionary forces (Robson (2001)). Alternatively, relative performance concerns may arise endogenously from strategic interaction, as in Basak and Makarov (2007).[45] The model that I develop is a parsimonious model of status concerns. An attractive feature of the model is that it nests both models from Section 6.4.

The model Each investment style in the mutual fund industry comprises of a continuum of mutual fund managers, in which each manager i, $i \in \mathcal{M}$, manages a fund of size A_{it} at time t. The total mass of managers is normalized to unity with a corresponding measure μ. The percentile rank of a fund of relative size a at time t is defined by:

$$\varrho_t(a) \equiv \mu\left(i \left| \frac{A_{it}}{\bar{A}_0 R_t^B} \leq a \right.\right), \tag{6.25}$$

[44] See for instance Robson (1992), Zou (1994), Zou (1995), Bakshi and Chen (1996), Carroll (2000), Chang, Hsieh, and Lai (2000), Cole, Mailath, and Postlewaite (2001), Goel and Thakor (2005), and Roussanov (2007). The latter two studies provide alternative motivations based on psychological and sociological foundations and evidence.

[45] This relates to the models of De Marzo, Kaniel, and Kremer (2004) and De Marzo, Kaniel, and Kremer (2007) in which investors compete for a scarce good. In the mutual fund industry, the scarce goods are fund flows that outside investors allocate across different funds.

where fund size is scaled by the median of the initial cross-sectional asset distribution,[46] $\bar{A}_0 \equiv \left\{\bar{A} \mid \mu\left(i \left| A_{i0} \leq \bar{A}\right.\right) = .5\right\}$, multiplied by the benchmark return, $R_t^B \equiv B_t/B_0$. I update the initial median fund size, $\bar{A}_0$, with the benchmark return to account for overall growth in assets under management during the year if the manager invests along with the pack. This implies that to improve status, the manager needs to deviate from the pack by increasing or decreasing passive risk, or by allocating capital to the active portfolio. I define $\bar{A}_t \equiv \bar{A}_0 R_t^B$. The manager's preferences are represented by:

$$\max_{(x_t)_{t\in[0,T]}} E_0\left[\frac{\eta}{1-\sigma_1} A_T^{1-\sigma_1} + (1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{1-\sigma_1}\,\varrho_T\left(\frac{A_T}{\bar{A}_T}\right)^{1-\sigma_2}\right], \tag{6.26}$$

with $\eta \in [0,1]$, $\sigma_1 > 1$, and $\mathcal{S}(x)$ as a sign function: $\mathcal{S}(x) = 1$ if $x \geq 0$ and $\mathcal{S}(x) = -1$ otherwise. The manager's utility is a weighted average of two terms with weights η and $(1-\eta)$, respectively. The first term summarizes the manager's preferences for assets under management. The second term captures status concerns. The term $\varrho_T\left(\frac{A_{iT}}{A_T}\right)$ represents the manager's position in the cross-sectional asset distribution. The curvature parameter σ_1 captures the manager's aversion to fluctuations in assets under management; σ_2 controls aversion to variation in fund status.

Several aspects deserve further discussion. First, the distribution function $\varrho_T(\cdot)$ is by definition bounded between zero and one. σ_2 can therefore be negative without inducing global convexities that would render the portfolio-choice problem ill-defined. In economic terms, managers with a strong desire to improve their status are identified by low, possibly even negative, values of σ_2. Managers who are concerned about variation in fund status have high values of σ_2. The desire to move up in the asset distribution can justify high levels of active risk-taking despite a lack of skill. Second, I assume that the preferences are separable in assets under management and in status concerns. This allows an interpretation in which the first part represents the current year's compensation and the second part captures the manager's value function over her remaining career prospects. Such career prospects presumably become less pressing when the manager's fund ranks higher in the cross-sectional asset distribution. In this interpretation, σ_2 measures the manager's career concerns. Third, the fund's rank is represented by the cumulative distribution function (CDF) of assets under management, $\varrho_T(\cdot)$, to simplify the interpretation. Theoretically, any increasing function of fund size can serve the same purpose, but the CDF captures the ease with which a manager can climb in the cross-sectional asset distribution. If the CDF is steep, a small increase in assets under management results in a substantial improvement in status. In contrast, a more

[46] Alternatively, Roussanov (2007) normalizes by the mean of the, in his application, wealth distribution. The median is empirically more stable than the mean as it curbs the impact of outliers. The resulting cross-sectional asset distribution is more stable over time.

dispersed asset distribution requires a more stellar performance to realize the same status improvement. Fourth, this model nests the models studied in Section 6.4. If $\eta = 0$, $\sigma_1 = 1$, and the asset distribution is uniform (that is, $\varrho_T(a) = a/C$, with C the upper-bound of the asset distribution), I recover the preferences in Section 6.4.1 with a coefficient of relative risk aversion σ_2. If $\sigma_2 = 1$, the model reduces to the preferences in Section 6.4.2 with a coefficient of relative risk aversion σ_1. Fifth, the second term is pre-multiplied by $\bar{A}_T^{1-\sigma_1}$. This implies that the preferences are invariant to changes in aggregate wealth (Roussanov (2007)). Sixth, I update the initial median fund size by the style benchmark return, $\bar{A}_T = \bar{A}_0 R_T^B$. Alternatively, I could use the return on the median fund. Using the style benchmark return has two advantages. First, the benchmark return is easy for managers to track and seems like the most visible target to beat. Second, the definition of managerial ability gets obfuscated when the manager can trade the median fund.[47] If all managers are skilled, the median fund return inherits this skill. This would imply that I estimate the manager's ability only in sofar as it surpasses the skill present in the initial median fund return. By using the style benchmark return to update the median fund size, the definition of skill is consistent with the first part of the paper and the extant literature.

Modeling the cross-sectional asset distribution As a first step towards analyzing the model empirically, I model the cross-sectional asset distribution, $\varrho_t(\cdot)$.[48] First, I assume that the cross-sectional asset distribution is log-normal with mean $\mu_{\varrho_t} \equiv E_t\left[\log\left(A_{it}/\bar{A}_t\right)\right]$ and standard deviation $\sigma_{\varrho t} \equiv \mathrm{Var}_t\left[\log\left(A_{it}/\bar{A}_t\right)\right]^{\frac{1}{2}}$, $\varrho_t(\cdot; \mu_{\varrho t}, \sigma_{\varrho t})$. Second, I assume that the asset distribution is stationary during the period $[0, T]$: $\mu_{\varrho T} = \mu_{\varrho 0}$ and $\sigma_{\varrho T} = \sigma_{\varrho 0}$. The first assumption is made for computational tractability. Because the main objective is to estimate the model for a large cross-section of managers, I need to impose some structure. The second assumption implies that the manager uses the asset distribution at the beginning of the year to make her assessment of status throughout the year. This assumption could be relaxed by allowing ϱ_T to be different from ϱ_0, but the manager would need to be able to hedge the risk of a shifting distribution to preserve market completeness.

I estimate the coefficients of the log-normal distribution (μ_{ϱ_0} and σ_{ϱ_0}) for each style and each year using the cross-section of funds at the beginning of the year. To estimate the

[47] One other alternative would be to update the median fund with the median fund return and restrict the asset menu to cash, the style benchmark, and the active portfolio. However, this renders the financial market to be dynamically incomplete.

[48] Note that the model endogenously generates a cross-sectional asset distribution, $\varrho_T(\cdot)$, given $\varrho_0(\cdot)$ and the cross-sectional distribution of managerial ability and risk preferences. For instance, Roussanov (2007) derives the stationary distribution that is consistent with the optimal policies of households in a life-cycle model. I am not merely interested in the stationary distribution, but also in the conditional distribution. In addition, I want to estimate ability and risk preferences for a large cross-section of managers. It is therefore computationally too intensive to impose the equilibrium condition as well. I therefore model the cross-sectional asset distribution directly.

cross-sectional asset distribution, I use all mutual funds in the CRSP data set. Clearly, it would be inappropriate to use only those funds for which I can identify the manager or management team. I test the appropriateness of the distributional assumption using the Jarque-Bera test of normality. The average $p-$value across all years ranges from 10.2% to 52.4% for the nine investment styles, which supports the normality assumption.[49]

Fund status, risk aversion, and risk-taking Fund size and the fund's position in the cross-sectional asset distribution play a key role in explaining risk-taking behavior. First, I discuss the link between relative fund size and risk aversion. Second, I show that the parameters σ_1 and σ_2 determine whether the manager adjusts either active or passive risk if risk aversion changes. The former (σ_1) controls passive risk-taking; σ_2 determines active risk-taking.

I first relate relative fund size and risk aversion. The status model is not homogenous in assets under management. To understand the implications for risk-taking, I define the coefficient of relative risk aversion, $RRA(a_t)$:

$$RRA(a_t) = -\frac{aJ_{t,aa}}{J_{t,a}}, \tag{6.27}$$

where J denotes the value function and subscripts partial derivatives.[50] If $t = T$, I obtain the Arrow-Pratt measure of risk aversion (Appendix 6.C provides further details). It is the weighted average of two terms, with $a_t \equiv A_t/\bar{A}_t$:

$$RRA\left(a_T\right) = \omega(a_T)\sigma_1 + (1-\omega(a_T))\left[\sigma_2 \frac{\varrho'\left(a_T\right)a_T}{\varrho\left(a_T\right)} - \frac{\varrho''\left(a_T\right)a_T}{\varrho'\left(a_T\right)}\right], \tag{6.28}$$

with weight:

$$\omega(a_T) \equiv \frac{\eta a_T^{-\sigma_1}}{\eta a_T^{-\sigma_1} + (1-\sigma_2)(1-\eta)\mathcal{S}(1-\sigma_2)\varrho\left(a_T\right)^{-\sigma_2}\varrho'\left(a_T\right)}. \tag{6.29}$$

For most empirically plausible parameter combinations, $\omega(a_T)$ goes from one to zero if a_T increases from zero to infinity. This implies that status concerns become more pressing if

[49] There exists an interesting parallel between modeling the cross-sectional asset distribution of mutual funds and the cross-sectional firm-size distribution (Luttmer (2007) and Lustig, Syverson, and Van Nieuwerburgh (2007)) or the size distribution (Gabaix (1999), Gabaix and Ioannides (2004), and Eeckhout (2004)). For the latter, it is still contested whether log-normality or a power law provides the correct description of the data (Eeckhout (2004)). An interesting open question is what generates the cross-sectional asset distribution in the mutual fund industry and what selection mechanisms are at play. Understanding the decision-making of fund managers is a first step in this direction.

[50] I define the coefficient of relative risk aversion based on the value function, which is the relevant measure of relative risk aversion for decision-making at time t. In the empirical section, I use $RRA(a_0)$ as the measure of risk aversion.

fund status increases. Equation (6.28) implies that the manager's coefficient of relative risk aversion combines three measures of relative risk aversion: (i) σ_1, (ii) $\sigma_2 \varrho'(a_T)a_T/\varrho(a_T)$, and (iii) $-\varrho''(a_T)a_T/\varrho'(a_T)$. The third measure is the relative risk aversion if preferences are linear in status only ($\eta = 0$ and $\sigma_2 = 0$). Figure 6.6 displays the three components if $\sigma_1 = 4$, $\sigma_2 = .5$, and $\eta = .0005$. The asset distribution is calibrated to the S&P 500. The horizontal axis plots $\log(a_T)$, the vertical axis the relative risk aversion. The first component (σ_1) is obviously invariant to size; the second component decreases in fund size, and the last component increases in fund size. For large funds, the third component always dominates the second component. The three components aggregate to the overall coefficient of relative risk aversion via the weight function ($\omega(a_T)$). For small funds, status concerns are irrelevant ($\omega(a_T) \simeq 1$) and risk aversion is solely governed by $\sigma_1 = 4$. By increasing the fund's assets under management, status concerns gradually become more important ($\omega(a_T) < 1$). As a result, the coefficient of relative risk aversion drops. In this region, the manager has a lot of scope to move up in the asset distribution by deviating from the pack. By moving further up in the cross-sectional asset distribution, the manager has little incentive to deviate from the pack for fear of losing her position in the distribution. Status concerns are key in this region ($\omega(a_T) \to 0$). The minimum risk aversion is attained around the 25-th percentile of the asset distribution. This implies that risk aversion increases in fund size for most funds.

The parameters σ_1 and σ_2 are key to understanding whether the manager modifies active or passive risk-taking if the coefficient of relative risk aversion changes. I show in Appendix 6.D that σ_1 controls passive risk-taking. If the manager decides to use passive risk to deviate from the pack, she can choose to increase or decrease the fund's beta. Either will lead to a tracking error relative to the average fund that has a unit beta. Appendix 6.D shows that the manager chooses to increase passive risk if $\sigma_1 < \lambda_B/\sigma_B$ and decreases passive risk if $\sigma_1 > \lambda_B/\sigma_B$. For $\sigma_1 = \lambda_B/\sigma_B$, the manager's passive risk-taking is insensitive to changes in the coefficient of relative risk aversion. Unlike passive risk-taking, active risk-taking always increases when σ_2 falls. This implies that σ_2 controls active risk-taking. The main problem with the models in Section 6.4 is that both active and passive risk-taking are proportional to the coefficient of relative risk aversion. An increase in the fund's beta goes hand-in-hand with an increase in active risk. The status model frees up this tight link.

I illustrate the role of σ_1, σ_2, and fund size by solving for the optimal initial allocation to the benchmark and the active portfolio. I set $\eta = .0005$ and $\lambda_A = .15$. Appendix 6.E.2 discusses the solution method. The results are presented in Table 6.8. The first four columns show the impact of σ_2. I set $a_0 = 1$ and $\sigma_1 = \lambda_B/\sigma_B$ so that x^B equals unity and is invariant to changes in σ_2. The main observation is that x^A is inversely related to σ_2, whereas relative risk aversion is positively related to σ_2. If σ_2 increases from -1 to 30, the optimal allocation drops from $x^A = 155\%$ to $x^A = 5\%$.

Columns five to twelve illustrate the role of σ_1 and the link between relative fund size and risk-taking. As before, I set $\sigma_2 = .5$. I consider the optimal allocation for different initial fund sizes. Columns five to eight consider the case in which $\sigma_1 = 3.75$ ($< \lambda_B/\sigma_B$), whereas the last four columns correspond to $\sigma_1 = 4.25$ ($> \lambda_B/\sigma_B$). First, risk aversion is (inversely) hump shaped as in Figure 6.6. If $\sigma_1 = 3.75$, the manager increases passive risk (x^B) as risk aversion drops, and decreases x^B if $\sigma_1 = 4.25$ for the same change in fund status. Second, the manager always increases active risk if risk aversion decreases. Note that x^A is virtually unaffected by changing σ_1, in particular for larger fund sizes. This implies that σ_1 controls passive risk-taking and σ_2 active risk-taking and provides a structural interpretation to the ideas of Litterman as iterated in the introduction. Risk aversion to passive risk translates into aversion to fluctuations in assets under management. Risk aversion to active risk translates into aversion to variation in fund status. In conclusion, fund status is for most funds positively related to risk aversion. How managers adjust risk-taking in response to a change in risk aversion is governed by σ_1 (passive risk) and σ_2 (active risk).

A key implication of the model is that managers of small funds will behave markedly different from managers controlling large funds. Small funds have more room to grow and to improve their status, which provides an incentive to deviate from the pack. The opposite is true for managers of large funds. As such, large funds will take less active risk and produce smaller alphas, consistent with empirical evidence on risk-taking and performance in relation to fund size.

Statistical identification The model contains four manager-specific parameters, $\Theta_A \equiv \{\sigma_1, \sigma_2, \eta, \lambda_A\}$. It turns out that η is weakly identified. I therefore calibrate η to a common value $\eta = .0005$. The previous section shows that this model can generate a wide variety of risk-return distributions.

6.8 Main empirical results

This section presents the empirical results for the status model.

The cross-sectional distribution of ability and risk aversion Table 6.9 summarizes the main estimation results by investment style, with the overall results across all styles in the bottom panel. First, all parameters are right-skewed, in particular σ_2. The coefficient of variation (the standard deviation normalized by the mean) is much larger for σ_2 than for σ_1 and λ_A. This points to substantial heterogeneity in status concerns. The dispersion in σ_1 is relatively small. This stems from the fact that σ_1 controls passive risk-taking and the empirical result that mutual fund betas display little dispersion.

The bottom panel shows that the average coefficient of relative risk aversion is estimated to be 5.16, with its median equal to 2.51 and a standard deviation of 7.69. The average manager has therefore a risk aversion coefficient that is slightly lower than the average household's risk aversion of 8.2 as estimated by Kimball, Sahm, and Shapiro (2007). It is appealing that mutual fund managers as a group are less conservative.

The average price of active risk is estimated to be .28, with a median equal to .14 and a standard deviation of .38. To put the estimates in perspective, I compare the model-implied estimates to the actual estimates of a performance regression (Appendix 6.A). Recall that the standard models in Section 6.4 cannot easily reproduce the coefficients of standard performance regressions. The model-implied estimates are computed as the average α, β, and σ_ε sampled at a monthly frequency. To gauge the similarity, I perform the following cross-sectional regression for, for instance, α:

$$\hat{\alpha}_i^{\text{Performance}} = \rho_0 + \rho_1 \hat{\alpha}_i^{\text{Status}} + u_i, \tag{6.30}$$

where $\hat{\alpha}_i^{\text{Performance}}$ is the estimate from a standard performance regression and $\hat{\alpha}_i^{\text{Status}}$ the estimate from the status model. The estimates following from the structural model are much sharper. I therefore use them as the right-hand side variables to mitigate the errors-in-variables bias due to estimation uncertainty. The resulting estimates read: for α, $\hat{\rho}_0 = -.00$ and $\hat{\rho}_1 = .99$ ($R^2 = 35.11\%$); for β, $\hat{\rho}_0 = -.00$ and $\hat{\rho}_1 = 1.00$ ($R^2 = 97.67\%$); for σ_ε, $\hat{\rho}_0 = -.00$ and $\hat{\rho}_1 = 1.04$ ($R^2 = 98.69\%$). In all cases it seems that the estimates are virtually unbiased ($\rho_0 = 0$ and $\rho_1 = 1$). The most striking result is the R-squared for the regression of mutual fund alphas.[51] The estimates from the structural model are three times more accurate and do a good job of reproducing the average moments of performance regressions. This is also illustrated in Figure 6.2, where the top panels provide the results for a standard performance regression and the bottom panels for the structural model. The left panels display the fund alphas before fees and expenses, the right panels are net of all expenses. The distribution of fund alphas following from the structural model is much less dispersed. This implies that the cross-sectional distribution of fund alphas following from performance regressions reflects predominantly estimation error and not heterogeneity in managerial ability or risk preferences.

Figure 6.3 displays a scatter plot of risk aversion (horizontal axis) and managerial ability (vertical axis) to analyze their interaction. The correlation between ability and risk aversion is 80.2%. A second-order polynomial fitted through this cloud shows that managerial ability

[51] This implies that in the reverse regression of α_i^{Model} on $\alpha_i^{\text{Performance}}$, $\hat{\rho}_1$ would be downward biased and, correspondingly, $\hat{\rho}_0$ upward biased. Indeed, the reverse regression results in $\hat{\rho}_0 = .013$ and $\hat{\rho}_1 = .36$, which motivates the regression specification in (6.30).

is increasing and concave in the coefficient of relative risk aversion. The last part of this section discusses potential mechanisms that can generate this positive relation.

There are also interesting differences across investment styles. I focus on large/value managers and small/growth managers. Figure 6.4 provides a standard kernel density estimate for risk aversion (left panel) and managerial ability (right panel). There are pronounced differences in the distribution of risk preferences for the two types of managers, despite the fact that the average risk aversion is very similar (5.66 for large/value and 5.49 for small/growth). Risk aversion is more evenly distributed for large/value managers, but it is more right-skewed for small/growth managers. The median risk aversion for the small/growth manager is 1.49, whereas the median large/value manager has a risk aversion of 3.95. Ability, by contrast, is considerably higher for small/growth managers on average, but their medians of .16 tie. This implies that there are more high-skilled managers in the small/growth investment style, which is reflected by the thicker tail of the distribution.

Heterogeneity in risk aversion and ability I relate the estimates of managerial ability and risk aversion to observable characteristics of managers and mutual funds using multiple cross-sectional regressions. The characteristics include total net assets, the manager's tenure, turnover, expenses, investment in common and preferred stocks, loads, 12B-1 fees, and the total net assets of the family. The results are presented in Table 6.10. I include dummies to absorb style-fixed effects and use standard errors that are robust to heteroscedasticity.

The dependent variables are expressed in logarithms and the independent variables are standardized. As such, the coefficients are to be interpreted as the percentage change for a one-standard deviation change in the characteristics. First, I find that skilled managers operate on smaller funds, consistent with Chen, Hong, Huang, and Kubik (2004), who document a negative relation between fund size and ability as measured by the fund's alpha. My structural model implies that a one-standard deviation increase in fund size leads to almost a 9% decrease in the price of active risk.[52] Second, managers with longer tenure periods are more skilled, which may be the outcome of selection based on skill or learning. A one-standard deviation increase in tenure increases the price of risk by 7%. Chevalier and Ellison (1999a) find the same sign for fund alphas as a measure for performance, but the effect is insignificant. Third, more skilled managers have higher levels of turnover[53] and have smaller stock holdings. Fourth, skilled managers charge higher expense ratios, consistent with Berk and Green (2004), but the effect is insignificant. Fifth, I find that more conservative managers manage larger funds, have smaller expense ratios, and allocate a smaller share of their

[52] See Edelen, Evans, and Kadlec (2007) and Pollet and Wilson (2007) for potential explanations for the relation between performance and fund size.

[53] Chen, Hong, Huang, and Kubik (2004) document a positive, but insignificant, relation between turnover and fund alphas. I find a significant relationship once corrected for heterogeneity in risk preferences.

capital to stocks. The relation between risk aversion and expenses is again consistent with Berk and Green (2004) because fund alphas and risk aversion are inversely related. Finally, note that there is considerable unobserved heterogeneity; the R-squared values are 13.0% for ability and 6.6% for risk aversion.

Testing competing models I study four models to describe mutual fund returns, of which three are structural (Section 6.4.1, 6.4.2, and 6.7) and one is reduced-form, namely performance regressions (Appendix 6.A). A valid question is whether the status model statistically improves the other three models. Since the relative-return model of Section 6.4.1 is nested only if the asset distribution is uniform (which is inconsistent with the data) and the performance regressions are non-nested, I use the test developed in Vuong (1989) to compare non-nested models (Appendix 6.F).

I perform the tests at the manager level for significance levels of 5% and 10%. Table 6.11 reports the averages across all managers in a particular style. The status model is favored if the average number of rejections exceeds the 5% or 10% significance level. The test results provide a clear ranking of the models. First, all three competing models are rejected in favor of the status model. It is important to note that the status model is also favored over performance regressions. This implies that the conditional distribution of the status model provides a better description of fund returns than performance regressions for which the conditional and unconditional distributions coincide. Therefore, the status model is able to capture important dynamics of mutual fund strategies that performance regressions cannot. Second, the rejection rates are highest for relative-return preferences, followed by preferences for assets under management. The reduced-form performance regression are most competitive, but are still rejected too often in favor of the status model. In conclusion, there is strong statistical support in favor of the status model.

The fraction of skilled managers Measuring mutual fund performance has been of great interest to both academics and practitioners. The recent view contends that there are only small number of fund managers who are able to recover their costs. For this group of managers, performance actually persists. Knowing which fraction of the managers possesses skill is key because most investors base their investment in active funds on this premise.[54] My approach provides a fresh look at this debate as the controversy stems from the large uncertainty surrounding the estimates of alpha.[55] Structural models of delegated

[54] Alternatively, investors may choose to invest in mutual funds for time considerations only. Mamaysky and Spiegel (2002) argue that investors can allocate their capital to mutual funds to complete the static investment opportunity set with dynamic strategies even if the dynamic strategies require no private information.

[55] Baks, Metrick, and Wachter (2001) address the question in a Bayesian way, while Kosowski, Timmermann, Wermers, and White (2006) use a bootstrap analysis to compute the correct, finite-sample distribution of the estimates. The former paper finds that even skeptical investors may allocate part of their capital to

management lead to sharper estimates of managerial ability as the cross-equation restrictions allow me to extract estimates of ability from the volatility of mutual fund returns.

Figure 6.2 displays the empirical distribution of mutual fund alphas following from performance regressions (top row) and the status model (bottom row). The left figures portray the fund alphas before costs, whereas the right figures subtract the fund's expenses. Before fees and expenses, the average alpha is 157bp for both performance regressions and for the status model. These numbers change to an average of 2bp after costs. This implies that the average alphas are zero, consistent with the prior literature.[56] However, the distribution of alphas from the structural model is much narrower. The fraction of alphas that exceed zero is therefore substantially smaller. For the performance regressions, 46.03% of the after-costs alphas exceed zero, while this number drops to 30.95% in the case of the status model.

The large number of managers that produce fund alphas that exceed fees and expenses reflects sampling error. I now study the managers who are able to reliably recuperate their costs. Statistical significance is determined using the asymptotic standard errors.[57] This is slightly more involved for the status model. To compute the standard errors for the status model, I first compute the fund's alpha as $\alpha(\Theta) = x_{A0}(\hat{\Theta}_A)\hat{\lambda}_A\tilde{\sigma}_A$, which I in turn average over all fund years.[58] I compute the standard errors by applying the delta theorem as $\sqrt{T}\left(\hat{\Theta}_A - \Theta_A\right) \rightarrow^d N\left(0, \Sigma_\Theta\right)$. For each manager, I test whether the after-costs alpha significantly exceeds zero at the 5% level. For the performance regressions, I can reject the null in 9.43% of the cases, while this number increases to 13.12% for the status model. This result is important because it shows that despite the fact that fewer managers recover their costs based on point estimates (31% instead of 46%), the increased efficiency of the estimator implies that more fund managers robustly display skill (13% instead of 9%). The fraction of skilled managers increases by almost 40%. Kosowski, Timmermann, Wermers, and White (2006) show that even a small number of skilled managers can be economically important, which underscores the economic relevance of this exercise.

To conclude, Table 6.12 depicts the fraction of managers that reliably recover their costs and expenses by investment style. Skilled managers are concentrated in the small/growth-oriented styles. For most investment styles, the structural estimation results in a more rosy

active management, and the latter paper estimates the fraction of skilled managers to be about 10%.

[56]Note that the risk-adjustment is only via the style benchmark. The results typically look somewhat worse if one also corrects using a four-factor model. Also, I require three years of return data to estimate the models, which introduces a survivorship bias. The results in Kosowski, Timmermann, Wermers, and White (2006) suggest, however, that this effect may be small.

[57]Alternatively, I could bootstrap the standard errors as in Kosowski, Timmermann, Wermers, and White (2006) to construct the finite-sample distribution of the test statistics. This would require frequent re-sampling and re-estimation of the structural model, which is computationally infeasible.

[58]The results are very similar if I sample the fund's alpha at a monthly frequency and subsequently average it over all fund years.

view of ability in the mutual fund industry.

Cross-sectional stability of ability and risk aversion A subset of fund managers in my dataset controls multiple funds belonging to different investment styles. This provides an opportunity to study the stability of ability and risk aversion estimates holding constant the economic environment. Obviously, it may be that a manager is more skilled in the large/value style than in small/growth or vice versa. Likewise, there may be disparity in fund sizes, which induces differences in risk aversion across styles. It would nevertheless be reassuring to detect a positive relationship across styles.

The sample contains 105 style matches for which I have at least three years of data. The resulting scatter plot of risk aversion and ability is displayed in Figure 6.7. The correlation in risk aversion estimates across styles equals 65.0%; it equals 32.9% for managerial ability. Both are significantly positive at the 1%-level. It implies that risk aversion estimates are stable across styles, which is important. Managerial ability is less stable, which may reflect that risk aversion is more an attribute of the manager, while ability is more asset-class specific.

Persistence of managerial ability I study whether measures of ability are useful to predict future performance.[59] I use either performance regressions or the status model to sort funds into quartiles based on a three-year selection period. I then form equally-weighted portfolios of funds in a particular quartile and compute the portfolio return and the corresponding equally-weighted benchmark return. I hold the portfolios for one year. This leads to a return series over the full sample period, which I use to compute the annualized information ratio. As an alternative, I include the momentum factor to correct for the returns on passive strategies. Figure 6.8 displays the main results for only benchmark returns (Panel A) or benchmark returns and the momentum factor (Panel B).

Regardless of the ranking procedure, the price of risk in the selection period relates positively to the information ratio in the out-of-sample period. The difference between both approaches is small, but performance regressions relate performance monotonically in both periods. By comparing Panel A and Panel B, I find that accounting for momentum trading reduces the information ratio, but the main pattern in rankings is unaffected. This implies that, based on this exercise, the structural model does not ameliorate the forecasting power of fund performance.

Note that I would ideally use the status model to measure performance in the out-of-sample period. This would be most consistent with the selection year and enhance efficiency.

[59] Brown and Goetzmann (1995), Elton, Gruber, and Blake (1996), Carhart (1997), Kosowski, Timmermann, Wermers, and White (2006) study the persistence of mutual fund performance.

In addition, funds take heterogeneous and time-varying amounts of active risk and passive risk. It is unclear how this affects standard performance regressions and the predictability results that I report. However, it is computationally too intensive to estimate the model for a large cross-section of fund managers over rolling samples.

Time series of relative risk aversion and expected returns The status model is not homogenous in assets under management. This implies that variation in both fund status and the cross-sectional asset distribution lead to variation in the coefficient of relative risk aversion. Both will move around the average coefficient of relative risk aversion across managers. The solid line in Figure 6.5 displays that average coefficient from 1992 to 2006. In recent equilibrium models featuring habit formation, time variation in risk aversion translates into time variation in risk premia (Campbell and Cochrane (1999)). It is therefore interesting to compare the resulting time series with the time series for expected returns, which is taken from Binsbergen and Koijen (2007). They use a present-value model to estimate the time series of expected returns and expected growth rates, which results in stronger predictors for future returns and dividend growth rates than standard predictive regressions. The dashed line corresponds to the time series of expected returns from 1992 to 2006. The two time series display a strong co-movement; their correlation equals 62.6%. This lends further credibility to the risk aversion estimates and its variation over time. I also compute the average price of active risk over the sample period (not reported). This average is very stable and varies in a range of only .05 over time.

Correlation risk aversion and managerial ability One empirical finding that is remarkably robust across all models is that risk aversion and managerial ability are positively correlated. Three potential mechanisms can generate this empirical regularity. First, it may simply be a genetic feature that skilled investors tend to be more conservative. Second, even if ability and risk aversion are uncorrelated in population, selection effects can lead to an increasing and concave relation between ability and managerial risk aversion, consistent with Figure 6.3. For expositional reasons, I focus on the model of Section 6.4.2, but the argument applies to all models. Consider an individual who can choose between a job in the mutual fund industry and a less risky job at a savings bank. For argument's sake, suppose the bank provides a known and constant income O_T at $t = T$. I assume that the manager decides which job to take based on one-period utilities, but the argument extends easily to a multi-period framework. As such, the manager compares the value function corresponding to the mutual fund industry ($A_0 = 1$):

$$J^{MF} = \frac{1}{1-\gamma} \exp\left((1-\gamma)r + \frac{1-\gamma}{2\gamma} \left(\lambda_A^2 + \lambda_B^2 \right) \right), \tag{6.31}$$

with the value function induced by the outside option $J^{OO} = \frac{1}{1-\gamma} O_T^{1-\gamma}$. The indifference locus reads:

$$\bar{\lambda}_A(\gamma) = \sqrt{(\log O_T - r)2\gamma - \lambda_B^2}. \tag{6.32}$$

Fund managers will opt into the industry only if $\lambda_A \geq \bar{\lambda}_A(\gamma)$. The right-hand side of (6.32) is increasing and concave in γ. Hence, even when ability and risk aversion are uncorrelated in population, selection effects may lead to the relation between ability and risk aversion. A third explanation would be that the status component in the utility index implicitly proxies for career concerns. Skilled managers may act more cautiously, realizing that they have more at stake than less skilled managers. The status component of the utility function can be interpreted as a value function or continuation utility. Consistent with the prediction that skilled managers are more status concerned, I find that the correlation between λ_A and σ_2 is positive and equals 57%. I show in Section 6.4.2 that a model with career concerns is unable to affect optimal policies for the relevant range of risk aversion. One reason why this model has so little bite is a peso-problem in measuring career concerns. If all managers avoid particular actions as they know this will induce demotion, then the model of demotion probabilities needs to extrapolate into this region and underestimate true career concerns faced by fund managers.

6.9 Optimal delegated investment management

In this section, I demonstrate that the joint distribution of risk preferences and managerial ability is a key input for the optimal allocation to active and passive management. I consider three types of investors who allocate capital to cash, a passive style index, and actively-managed mutual funds.[60] They differ in how they account for heterogeneity. The first investor optimally accounts for heterogeneity and uses the structural model of Section 6.7. The second investor ignores heterogeneity and uses sample average values of the parameters that govern ability and risk preferences. The third investor accounts for heterogeneity, but uses performance regressions to characterize heterogeneity. I quantify the economic importance of heterogeneity by measuring the utility costs that are induced by these sub-optimal strategies.

[60] I focus in the section on the problem of delegated instead of decentralized investment management. The current approach can easily be extended to study the decentralized problem with different investment styles (Binsbergen, Brandt, and Koijen (2007)).

The economic importance of heterogeneity The investor is assumed to have power-utility preferences with a coefficient of relative risk aversion γ_I:

$$\max_{\pi} E_0\left[\frac{1}{1-\gamma_I}W_T^{1-\gamma_I}\right], \tag{6.33}$$

where W_T denotes time-T wealth. For tractability, I assume that the investor implements a constant-proportions strategy, π,[61] so that wealth evolves as:

$$\mathrm{d}W_t = W_t\left(r + \pi'\Sigma_t\Lambda\right)\mathrm{d}t + W_t\pi'\Sigma_t\mathrm{d}Z_t, \tag{6.34}$$

with $\Sigma_t \equiv (\sigma_P, \Sigma' x_t(\Theta_A))'$ and $x_t(\Theta_A)$ indicates the manager's optimal (dynamic) strategy. The returns produced by the active manager depend on her ability and risk preferences, summarized in Θ_A. A key feature of (6.33) is that the expectation operator not only integrates out financial risks, but also uncertainty about the manager's type, that is, her ability and risk preferences. Denote the return distribution conditional on the manager's parameters by $f_{R|\Theta_A}(R_T \mid \Theta_A)$, and the distribution over the managers' parameters by $f_{\Theta_A}(\Theta_A)$. Their joint distribution is denoted by $f_{R,\Theta_A}(R_T, \Theta_A) \equiv f_{R|\Theta_A}(R_T \mid \Theta_A) \cdot f_{\Theta_A}(\Theta_A)$. The problem in (6.33) can then be reformulated as:

$$\begin{aligned}\max_{\pi} E_0\left[\frac{1}{1-\gamma_I}W_T^{1-\gamma_I}\right] &= \max_{\pi}\int_{\Theta_A}\left[\int_R \frac{1}{1-\gamma_I}W_T^{1-\gamma_I} f_{R|\Theta_A}(R_T \mid \Theta_A)\,\mathrm{d}R\right] f_{\Theta_A}(\Theta_A)\mathrm{d}\Theta_A \\ &= \max_{\pi}\int_{\Theta_A} J(\pi, T, \Theta_A) f_{\Theta_A}(\Theta_A)\mathrm{d}\Theta_A \equiv \max_{\pi}\tilde{J}(\pi, T),\end{aligned} \tag{6.35}$$

with $J(\pi, T, \Theta_A)$ denoting the value function corresponding to the strategy π when the manager's parameters are given by Θ_A. To quantify the utility costs induced by ignoring heterogeneity in ability and preferences, I compare the optimal solution to (6.35), denoted by π_1, to the optimal strategy of an investor who employs the average parameters instead, denoted by $\pi_2 \equiv \arg\max_\pi J(\pi, T, E[\Theta_A])$. The utility costs are computed as the reduction in certainty-equivalent wealth:

$$CEQ(\pi_1, \pi_2) \equiv \left(\tilde{J}(\pi_2, T)/\tilde{J}(\pi_1, T)\right)^{\frac{1}{1-\gamma_I}} - 1. \tag{6.36}$$

Appendix 6.G provides details on how to compute $CEQ(\pi_1, \pi_2)$. If the investor uses performance regressions, the optimal strategy can be determined along the same lines. The investor

[61] Since the investor implements a constant-proportions strategy, wealth (W_I) is ensured to be non-negative. The investor never shorts the mutual fund, which carries a non-negative price of risk. Aggressive investors may short the index or borrow to invest in mutual funds and the style index, which is feasible in real life.

allocates her capital to the style benchmark and the managed portfolio with dynamics:

$$\mathrm{d}A_t = A_t \left(r + \beta\sigma_P\lambda_P + \alpha\right)\mathrm{d}t + A_t\beta\sigma_P\mathrm{d}Z_t^P + A_t\sigma_\varepsilon\mathrm{d}Z_t^A, \tag{6.37}$$

where the coefficients $\Theta_A = (\alpha, \beta, \sigma_\varepsilon)$ are drawn from the empirical distribution. In this case, heterogeneity is reflected by the trivariate empirical distribution of Θ_A, which the manager integrates out as in (6.35). Further details are provided in Appendix 6.G.

Empirical results Panel A of Figure 6.9 displays the optimal allocation to the style benchmark (left) and mutual funds (right) for the three investors in the large/value style. The horizontal axis displays the coefficient of relative risk aversion ranging from one to 10, the vertical axis the optimal allocation. First, the investor who ignores heterogeneity in ability and risk aversion (blue solid line) invests more aggressively in mutual funds than the optimal strategy (green dashed line). Heterogeneity acts as a source of background risk or Bayesian parameter uncertainty, which increases the effective risk aversion of the investor (Gollier and Pratt (1996)). Second, if the investor uses performance regressions, she becomes overly conservative. The estimated cross-sectional distribution is a convolution of true heterogeneity and estimation error. For performance regressions (red dotted line), the latter component dominates and the manager overestimates heterogeneity. The utility costs corresponding to both sub-optimal strategies are depicted in Panel B for the large/value style (left) and small/growth style (right). First, the utility costs are economically large, in particular for the small/growth style. For the strategy that ignores heterogeneity (blue solid line), the utility costs can be as a high as 380bp per year, while the costs peak above 50bp for the large/value style. Second, the use of performance regressions somewhat mitigates the costs (red dotted line). For the small/growth style, the maximum costs drop to 130bp. In sum, the utility costs are substantial, suggesting that it is crucial for investors to take into account the heterogeneity in ability and risk preferences of their mutual fund managers.

6.10 Conclusions

I use structural models of delegated portfolio management to recover the cross-sectional distribution of managerial ability and risk aversion. I develop a new likelihood-based estimation procedure to analyze such models empirically. By imposing the cross-equation restrictions that are implied by the structural models, I show that both managerial ability and risk preference parameters can be estimated from the volatility instead of the mean of fund returns. As such, I obtain sharp estimates of managerial ability, an issue that has plagued the performance literature ever since Jensen (1968). I find that 31% of the managers have positive

alphas after costs. Once sampling uncertainty is taken into account, this number drops to 13%.

Two standard models of delegated portfolio management result in economically implausible estimates of either managerial ability or risk aversion. Therefore, I develop a new model that imputes a concern for the relative position in the cross-sectional asset distribution into the preferences of the manager. I find that this model describes fund returns better than the other structural models and reduced-form performance regressions. The resulting estimates of managerial ability and risk aversion are plausible.

The main empirical results can be summarized as follows. First, risk aversion and managerial ability are both right-skewed, and there is more heterogeneity in risk preferences than in ability. Second, risk aversion and managerial ability are positively related. Skilled managers are more cautious. I show that this result can be explained by selection arguments or career concerns. Third, only a small fraction of the cross-sectional variation can be related to observable characteristics, which points to considerable unobserved heterogeneity. Fourth, the model endogenously generates time variation in risk aversion. I find that this time variation strongly co-moves with the equity risk premium; their correlation is 62%.

My results can be extended in several directions. First, the methodology that I develop may be applied to a range of different problems that use martingale techniques. Such applications include dynamic models with strategic interaction (Basak and Makarov (2007)), ability and preferences in the hedge fund industry (Panageas and Westerfield (2007)), and dynamic corporate finance models. Second, learning about managerial ability plays a key role in many theoretical mutual fund models (for instance, Berk and Green (2004) and Dangl, Wu, and Zechner (2007)). The approach in this paper implies that the individual investor's learning mechanism is much more efficient if the investment problem of the manager is taken into account. It seems therefore interesting to revisit the role of learning when explaining phenomena in mutual fund markets using the approach advocated in this paper. Third, recent models of consumption-based asset pricing use the household's Euler condition to price the assets. However, most capital invested in financial markets flows through the hands of delegated portfolio managers. Several recent studies show that the manager can become the inframarginal agent that prices the assets (for instance He and Krishnamurthy (2006)). If this is the case, deepening our understanding of the preferences of mutual fund managers is an important component of a better understanding of asset prices. The pronounced comovement between risk aversion and risk premia I find suggests that there is merit to this conjecture. The results in this paper provide a first step in modeling the preferences of the managers that decide upon the optimal asset allocation on behalf of most households. Explicitly incorporating the intermediation sector in consumption-based asset pricing models

is left for future research. Finally, it is interesting to explicitly model the manager's private information as in Liu, Peleg, and Subrahmanyam (2007). Information on the manager's returns and portfolio holdings[62] can then be used to extract information about the manager's quality of private information and risk preferences, which builds upon recent work of Cohen, Coval, and Pástor (2005), Wermers, Yao, and Zhao (2007), and Yuan (2007).

6.A Performance regressions in continuous time

This appendix summarizes performance regressions in a continuous-time framework. Nielsen and Vassalou (2004) discuss the link between continuous-time and discrete-time performance measures. I focus on the case with one style benchmark, but extensions to multi-factor benchmark models are trivial. In continuous time, the standard performance regression reads:

$$\frac{\mathrm{d}A_t}{A_t} - r\mathrm{d}t = \alpha\mathrm{d}t + \beta\left(\frac{\mathrm{d}S_t^B}{S_t^B} - r\mathrm{d}t\right) + \sigma_\varepsilon \mathrm{d}Z_t^A. \tag{6.38}$$

This implies that the dynamics of assets under management satisfy:

$$\frac{\mathrm{d}A_t}{A_t} = (r + \alpha + \beta\sigma_B\lambda_B)\,\mathrm{d}t + \beta\sigma_B\mathrm{d}Z_t^B + \sigma_\varepsilon\mathrm{d}Z_t^A. \tag{6.39}$$

For comparability with the structural models, I perform a two-step procedure in which I compute the likelihood of fund returns, $r^{A,\kappa\times T}$, conditional on benchmark returns, $r^{B,\kappa\times T}$, and the passive parameters, $\hat{\Theta}_B$, that are estimated in the first step. κ denotes the number of fund years available for a particular manager-fund combination. I define $\Theta_C \equiv \{\alpha, \beta, \sigma_\varepsilon\}$.

The performance parameters Θ_C are estimated by maximizing the log-likelihood:

$$\max_{\Theta_C} \mathcal{L}\left(r^{A,\kappa\times T} \mid r^{B,\kappa\times T}; \Theta_C, \hat{\Theta}_B\right) = \max_{\Theta_C} \sum_{t=h}^{\kappa T/h} \ell\left(r_t^A \mid r_t^B; \Theta_C, \hat{\Theta}_B\right). \tag{6.40}$$

Given the log-normal structure of the financial market in Section 6.3, the joint dynamics of the passive return and the mutual fund return are given by:

$$r_t^B = \left(\bar{r} + \sigma_B\lambda_B - \frac{1}{2}\sigma_B^2\right)h + \sigma_B\Delta Z_t^B, \tag{6.41}$$

$$r_t^A = \left(\bar{r} + \alpha + \beta\sigma_B\lambda_B - \frac{1}{2}\beta^2\sigma_B^2 - \frac{1}{2}\sigma_\varepsilon^2\right)h + \beta\sigma_B\Delta Z_t^B + \sigma_\varepsilon\Delta Z_t^A, \tag{6.42}$$

with $h = 1/12$ because the parameters are expressed in annual terms, $\bar{r}$ the average 1-month T-bill rate over the relevant year, $\Delta y_t \equiv y_t - y_{t-h}$, and:

$$\begin{pmatrix} \Delta Z_t^B \\ \Delta Z_t^A \end{pmatrix} \sim N\left(0_{2\times1}, hI_{2\times2}\right). \tag{6.43}$$

It therefore holds:

$$r_t^A \mid r_t^B \sim N\left(\mu_t, \sigma^2\right), \tag{6.44}$$

with:

$$\mu_t \equiv \left(\bar{r} + \alpha + \beta\sigma_B\lambda_B - \frac{1}{2}\beta^2\sigma_B^2 - \frac{1}{2}\sigma_\varepsilon^2\right)h + \beta\left(r_t^B - \left(\bar{r} + \sigma_B\lambda_B - \frac{1}{2}\sigma_B^2\right)h\right), \tag{6.45}$$

$$\sigma^2 \equiv \sigma_\varepsilon^2 h, \tag{6.46}$$

[62] Dybvig and Rogers (1997) propose a simple estimator of preference parameters based on holdings data.

which results in the log-likelihood in (6.40).

6.B Career concerns and fund flows

This appendix extends the model in Section 6.4.2 to allow for career concerns and external fund flows. The model closely follows Chapman and Xu (2007). Section 6.B.1 summarizes the model, while Section 6.B.2 provides further details on its calibration. Section 6.B.3 derives the Bellman equation and demonstrates its homogeneity in assets under management. Section 6.B.4 briefly summarizes the numerical procedure. The optimal strategies and results are discussed in Section 6.B.5. Time is expressed in months in this section to simplify notation.

6.B.1 The model

The dynamics of assets under management reads $A_t = \theta_t A_{t-3}$, with θ_t:

$$\theta_t \equiv \begin{cases} R_t^A \exp\left(F_{t-3}(z_{t-3}) + \varepsilon_t^F\right) & \text{, w.p.} \quad 1 - q^P_{t|t-3}(z_{t-3}, Age_{t-3}) - q^D_{t|t-3}(z_{t-3}, Age_{t-3}) \\ \nu_P & \text{, w.p.} \quad q^P_{t|t-3}(z_{t-3}, Age_{t-3}) \\ \nu_D & \text{, w.p.} \quad q^D_{t|t-3}(z_{t-3}, Age_{t-3}) \end{cases},$$

with $R_t^A \equiv A_t/A_{t-3}$ and $q^P_{t|t-3}$ ($q^D_{t|t-3}$) the probability that the manager will be promoted (demoted) at time t conditional upon the information at time $t-3$. The change in assets under management in case of promotion (demotion) is denoted by $\nu_P > 1$ ($\nu_D < 1$). F_{t-3} denotes the expected fund flow and $\varepsilon_t^F \sim N\left(0, \sigma_F^2\right)$ is the idiosyncratic risk present in fund flows.[63] Fund flows and promotion/demotion probabilities depend on past fund performance via z_t, which evolves as:

$$z_t = \rho_0 z_{t-3} + \rho_1 \left(R_t^A - R_t^B\right), \tag{6.47}$$

and forms a weighted average of past relative performance. Promotion and demotion probabilities furthermore depend on the number of years that the manager is active in the mutual fund industry, Age_t. Appendix 6.B.2 describes the exact functional forms and calibration in full detail, derives the value function, and provides further details on the numerical method. The decision frequency is quarterly and I assume that the manager follows a constant-proportions strategy at intermediate points in time.[64]

6.B.2 Model specification and calibration details

Fund flows are modeled as a third-order polynomial in past performance:

$$F_t = \delta_0 + \sum_{i=1}^{3} \delta_i \cdot (z_t)^i,$$

of which the parameters are given in the Table 6.1. It also reports the idiosyncratic volatility of fund flows (σ_F) and the increase (decrease) in assets under management, ν_P (ν_D), in case of promotion (demotion).

The promotion and demotion probabilities are represented by a multinomial logit model:

$$q^P_{t|t-h} = \frac{\exp(\varphi'_P x_t)}{1 + \exp(\varphi'_P x_t) + \exp(\varphi'_D x_t)},\ q^D_{t|t-h} = \frac{\exp(\varphi'_D x_t)}{1 + \exp(\varphi'_P x_t) + \exp(\varphi'_D x_t)},$$

with $x_t \equiv (z_t, Age_t)'$. The parameters that describe the promotion and demotion probabilities are depicted in Table 6.2. The variable Age_t indicates the period that the manager is active in the industry and is used to compute the dummy variables in Table 6.2. The performance variable z_t evolves according to (6.47). The

[63] Uncertainty in fund flows is assumed to be independent of the other financial risks.

[64] This assumption follows Campbell and Viceira (1999) and Campbell, Chan, and Viceira (2003). Under this assumption, the investor can hold long and short positions without rendering the strategy inadmissible.

	δ_0	δ_1	δ_2	δ_3
Value	0.0135	0.0928	-0.0031	-0.0371
Growth	0.0142	0.1389	0.0411	-0.0703

ν_D	ν_P	σ_F
0.423	1.72	0.13

Table 6.1: **Model parameters**
The table lists the parameters for the model of fund flows in Appendix 6.B. δ_i, $i = 0, \ldots, 3$, describe how fund flows depend on past performance (see (6.B.2)) and σ_F is the idiosyncratic risk in fund flows. The table also displays the (proportional) reduction in assets under management in case of demotion (ν_D) and the (proportional) increase in case of promotion (ν_P). The estimates are taken from Chapman, Evans, and Xu (2007).

	Variables contained in x_t					
Promotion	z_t	$I_{(\text{Tenure}\leq 3)}$	$I_{(\text{Tenure}\in(3,7])}$	$I_{(\text{Tenure}>7)}$	$z_t I_{(\text{Tenure}\leq 3)}$	$z_t I_{(\text{Tenure}\in(3,7])}$
Value	0.9327	-4.8866	-4.6177	-4.5313	0.0244	-0.2620
Growth	0.9118	-4.8004	-4.6615	-4.6746	0.2188	0.0482
Demotion						
Value	-0.6682	-3.4784	-3.5454	-3.8634	-0.0884	-0.2415
Growth	-0.5930	-3.7300	-3.7186	-3.8829	0.0331	-0.1348

Table 6.2: **Model parameters**
The table lists the parameters for the model of managerial promotion and demotion in Appendix 6.B. It contains the parameters of the multi-nominal logit model. The estimates are taken from Chapman, Evans, and Xu (2007).

parameters for the value manager equal: $\rho_0 = 0.51$ and $\rho_1 = 0.178$; for growth managers: $\rho_0 = 0.59053$ and $\rho_1 = 0.15309$.

6.B.3 Homogeneity of the value function

The manager's problem is given by:

$$\max_{\{x_0, x_3, x_6, x_9\}} E_0 \left[\frac{1}{1-\gamma} A_T^{1-\gamma} \right]. \tag{6.48}$$

Define the manager's value function as:

$$J(A_t, z_t, t) = \max_{\{x_t, \ldots, x_{T-3}\}} E_t \left[\frac{1}{1-\gamma} A_T^{1-\gamma} \right], \tag{6.49}$$

with $J(A_T, z_T, T) \equiv \frac{A_T^{1-\gamma}}{1-\gamma}$. The value function satisfies the Bellman equation:

$$J(A_t, z_t, t) = \max_{x_t} E_t \left[J(A_{t+3}, z_{t+3}, t+3) \right]. \tag{6.50}$$

I show that value function has the property:

$$\begin{aligned} J(A_t, z_t, t) &= A_t^{1-\gamma} J(1, z_t, t) \\ &= A_t^{1-\gamma} \tilde{J}(z_t, t), \end{aligned} \tag{6.51}$$

with $\tilde{J}(z_t, t) \equiv J(1, z_t, t)$. The proof is by induction. At $t = T$, the property trivially holds. Suppose that (6.51) also holds for s, $t < s \leq T$, then it follows:[65]

$$\begin{aligned} J(A_t, z_t, t) &= \max_{x_t} E_t \left[J(A_{t+3}, z_{t+3}, t+3) \right] \\ &= \max_{x_t} E_t \left[A_{t+3} \tilde{J}(z_{t+3}, t+3) \right] \\ &= \max_{x_t} A_t^{1-\gamma} \left[\begin{array}{l} q^P_{t+3|t} \nu_P^{1-\gamma} E_t \left[\tilde{J}(z_{t+3}, t+3) \right] + q^D_{t+3|t} \nu_D^{1-\gamma} E_t \left[\tilde{J}(z_{t+3}, t+3) \right] + \\ \left(1 - q^P_{t+3|t} - q^D_{t+3|t} \right) E_t \left[R_{t+3}^{A(1-\gamma)} \exp\left(F_t(1-\gamma) + \varepsilon^F_{t+3}(1-\gamma) \right) \tilde{J}(z_{t+3}, t+3) \right] \end{array} \right] \\ &= \max_{x_t} A_t^{1-\gamma} \tilde{J}(z_t, t), \end{aligned} \tag{6.52}$$

which establishes the homogeneity of the value function.

6.B.4 Numerical procedure

The optimal allocation to the style benchmark and active portfolio are determined by means of numerical dynamic programming. Since the model is specified at a quarterly frequency, the manager needs to make four investment decisions per annum, conditional on prior performance, $x_t^\star(z_t)$. The optimal policies are determined using the Bellman equation derived in the previous section. The model features three shocks; two for the financial market and one for idiosyncratic fund flows. However, the latter shock only enters the value function in case the manager is not demoted nor promoted:

$$\left(1 - q^P_{t+3|t} - q^D_{t+3|t} \right) E_t \left[R_{t+3}^{A(1-\gamma)} \exp\left(F_t(1-\gamma) + \varepsilon^F_{t+3}(1-\gamma) \right) \tilde{J}(z_{t+3}, t+3) \right],$$

see (6.52). As a result, this shock can be integrated out analytically:

$$\begin{aligned} &\left(1 - q^P_{t+3|t} - q^D_{t+3|t} \right) E_t \left[R_{t+3}^{A(1-\gamma)} \exp\left(F_t(1-\gamma) + \varepsilon^F_{t+3}(1-\gamma) \right) \tilde{J}(z_{t+3}, t+3) \right] = \\ &\left(1 - q^P_{t+3|t} - q^D_{t+3|t} \right) \exp\left(F_t(1-\gamma) + \frac{1}{2}(1-\gamma)^2 \sigma_F^2 \right) E_t \left[R_{t+3}^{A(1-\gamma)} \tilde{J}(z_{t+3}, t+3) \right], \end{aligned}$$

which implies that only two shocks are left. I use bivariate Gaussian quadrature to compute all expectations that arise (Tauchen and Hussey (1991)). For the performance variable z_t, I form an equally-spaced grid on $[-0.90, 0.90]$ with steps of size 0.05. I solve for the optimal strategy and the implied value function at each of the grid points. Refining the step sizes does not change the results. The value function in between grid points is interpolated via cubic-spline interpolation.

6.B.5 Optimal strategies

Figure 6.1 depicts the optimal investment strategy of a large/value manager that works between 3-7 years in the mutual fund industry.[66] The top figure displays the manager's holdings of the benchmark asset and the bottom figure the optimal allocation to the active portfolio. The axes correspond to the manager's coefficient of relative risk aversion (γ) and prior relative performance (z_t). The downward sloping plane corresponds to the problem in which there are no incentives ($\theta_t = R_t^A$), while the non-monotone plane describes the optimal solution to the full-fledged model. Three aspects are worth mentioning. First, the optimal allocations are non-monotone in risk aversion. The intuition is as follows. As (6.52) shows, the value function has three

[65] For notational convenience, I omit the arguments of the promotion/demotion probabilities and expected fund flows.

[66] The results for the small/growth managers and different tenure stages are qualitatively similar.

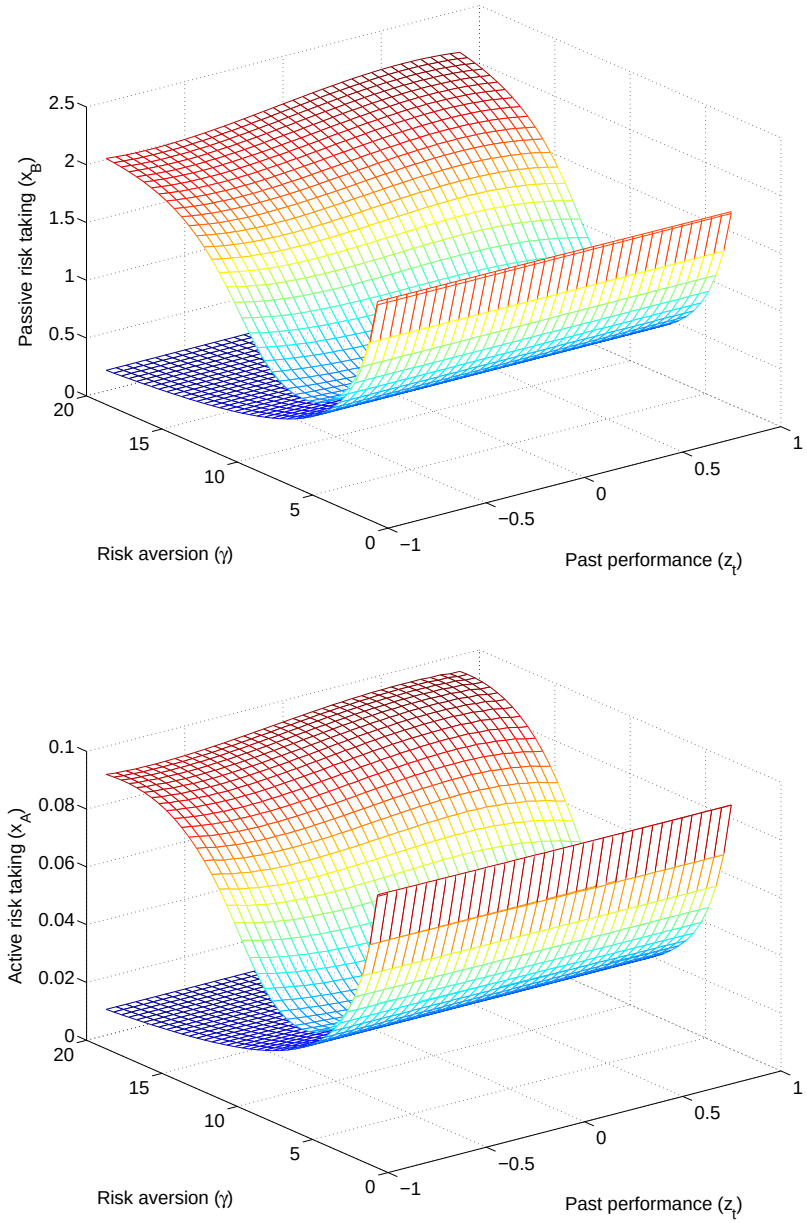


Figure 6.1: Optimal investment strategy in the presence of incentives
The figure displays the optimal allocation to the style benchmark (top panel) and the active portfolio (bottom panel) for different values of the coefficient of relative risk aversion (γ) and past performance (z_t). z_t is defined in (6.47). The monotone planes correspond to the optimal strategy of the model in Section 6.4.2. The non-monotone planes display the optimal strategies in the model of Appendix 6.B. In this model, assets under management change due to fund returns, performance-sensitive fund flows, and promotion and demotion. The figure depicts the optimal investment strategy of a large/value manager that works between 3-7 years in the mutual fund industry.

components corresponding to promotion, demotion, and no career change. The demotion event results in a drop in assets in which case the value function is pre-multiplied by $\nu_D^{1-\gamma}$. Since $\nu_D < 1$, this number turns large when γ increases, and the manager effectively minimizes the demotion probability, which is virtually linear in performance. As a result, the manager acts as if she is more aggressive. Second, for low levels

of risk aversion, incentives do not affect the optimal investment strategy. This region turns out to be key however. Mutual funds have betas with respect to their style benchmarks that are close to one and display relatively little dispersion. In this model, this can be reconciled by either a risk aversion level that is close to λ_B/σ_B in case incentives have no bite and the manager acts as an asset-only investor. Alternatively, the manager's risk aversion is such that it exactly balances the demotion probability in the increasing region in the direction of risk aversion. The latter case can however be ruled out for different economic reasons. It is well known that older managers take more risk (Chevalier and Ellison (1997)), which in fact motivates Chapman and Xu (2007) to study the impact of career concerns on risk-taking. Older managers are less likely to be demoted, meaning that the increase in risk-taking for more conservative managers becomes less pronounced. Hence, the model predicts that older managers to take on less risk, which clearly is at odds with the data. Therefore, I conclude that incentives, once calibrated to observed career changes and fund flows, hardly affect the optimal policies in the relevant region of risk aversion. As such, I study the reduced-form case with $\theta_t = R_{At}$ from now on for this model. Third, in the direction of past relative performance, there is a slight increase in risk-taking. This is a consequence of the convexities in fund flows. Indeed, if F_t is zero, the slope in this direction is nearly zero. The economic significance is small however, which resonates with the findings of Chapman and Xu (2007).

6.C Relative risk aversion in the status model

The manager's preferences are given by:

$$E_0\left[u\left(A_T,\bar{A}_T\right)\right] = E_0\left[\frac{\eta}{1-\sigma_1}A_T^{1-\sigma_1} + (1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{1-\sigma_1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{1-\sigma_2}\right]. \tag{6.53}$$

The coefficient of relative risk aversion reads:

$$RRA\left(A_T,\bar{A}_T\right) \equiv -\frac{A_T u^{(2,0)}\left(A_T,\bar{A}_T\right)}{u^{(1,0)}\left(A_T,\bar{A}_T\right)}, \tag{6.54}$$

with $u^{(i,j)}$ denoting the i-th derivative with respect to A_T and the j-th derivative to $\bar{A}_T$. The required derivatives are given by:

$$\begin{aligned}
u^{(1,0)}\left(A_T,\bar{A}_T\right) &= \eta A_T^{-\sigma_1} + (1-\sigma_2)\,(1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2}\varrho'\left(\frac{A_T}{\bar{A}_T}\right) && (6.55)\\
u^{(2,0)}\left(A_T,\bar{A}_T\right) &= -\sigma_1\eta A_T^{-\sigma_1-1} - \sigma_2\,(1-\sigma_2)\,(1-\eta)\,S\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1-1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2-1}\left[\varrho'\left(\frac{A_T}{\bar{A}_T}\right)\right]^2\\
&\quad + (1-\sigma_2)\,(1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1-1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2}\varrho''\left(\frac{A_T}{\bar{A}_T}\right)\\
&= -\sigma_1\eta A_T^{-\sigma_1-1} - && (6.56)\\
&\quad (1-\sigma_2)\,(1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1-1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2}\left[-\varrho''\left(\frac{A_T}{\bar{A}_T}\right) + \sigma_2\frac{\left[\varrho'\left(\frac{A_T}{\bar{A}_T}\right)\right]^2}{\varrho\left(\frac{A_T}{\bar{A}_T}\right)}\right],
\end{aligned}$$

which implies that the Arrow-Pratt measure of relative risk aversion reads:

$$\begin{aligned}
RRA\left(A_T,\bar{A}_T\right) &= \frac{\sigma_1\eta A_T^{-\sigma_1} + (1-\sigma_2)\,(1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2}\frac{A_T}{\bar{A}_T}\left[-\varrho''\left(\frac{A_T}{\bar{A}_T}\right) + \sigma_2\frac{\left[\varrho'\left(\frac{A_T}{\bar{A}_T}\right)\right]^2}{\varrho\left(\frac{A_T}{\bar{A}_T}\right)}\right]}{\eta A_T^{-\sigma_1} + (1-\sigma_2)\,(1-\eta)\,\mathcal{S}\,(1-\sigma_2)\,\bar{A}_T^{-\sigma_1}\varrho\left(\frac{A_T}{\bar{A}_T}\right)^{-\sigma_2}\varrho'\left(\frac{A_T}{\bar{A}_T}\right)}\\
&= \omega(a_T)\sigma_1 + (1-\omega(a_T))\left[\sigma_2\frac{\varrho'\left(a_T\right)a_T}{\varrho\left(a_T\right)} - \frac{\varrho''\left(a_T\right)a_T}{\varrho'\left(a_T\right)}\right], && (6.57)
\end{aligned}$$

with:

$$\omega(a_T) \equiv \frac{\eta a_T^{-\sigma_1}}{\eta a_T^{-\sigma_1} + (1-\sigma_2)(1-\eta)\mathcal{S}(1-\sigma_2)\varrho(a_T)^{-\sigma_2}\varrho'(a_T)}. \tag{6.58}$$

Note that both $\varrho > 0$ and $\varrho' > 0$, which implies $\omega(a_T) \in [0,1]$.

6.D The role of σ_1 in passive risk-taking

I show in this appendix that σ_1 controls passive risk taking in the status model of Section 6.7. I solve for the optimal strategy using the martingale method, which relates to Basak, Pavlova, and Shapiro (2007b). In the martingale approach, I first solve for the optimal terminal asset level. The optimal investment strategy is then given by the strategy that replicates this terminal claim.

Using the homogeneity property of the value function, the manager's problem can be reformulated as:

$$\max_{R_T^A \geq 0} E_0\left[\frac{\eta}{1-\sigma_1}\left(a_0 R_T^A\right)^{1-\sigma_1} + (1-\eta)\mathcal{S}(1-\sigma_2)R_T^{B(1-\sigma_1)}\varrho_T\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_2}\right], \tag{6.59}$$

with $R_t^A \equiv A_t/A_0$. The optimization is subject to the static budget constraint (recall that $\varphi_0 = 1$):

$$E_0\left[R_T^A\varphi_T\right] = 1. \tag{6.60}$$

The corresponding Lagrangian reads, with ξ denoting the Lagrange parameter:

$$\begin{aligned}
\mathcal{Z}(R_T^A, R_T^B, \xi) &= \frac{\eta}{1-\sigma_1}\left(a_0 R_T^A\right)^{1-\sigma_1} + (1-\eta)\mathcal{S}(1-\sigma_2)R_T^{B(1-\sigma_1)}\varrho_T\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_2} - \xi\varphi_T R_T^A \\
&\propto \frac{\eta}{1-\sigma_1}\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_1} + (1-\eta)\mathcal{S}(1-\sigma_2)\varrho_T\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_2} - \xi\varphi_T\frac{R_T^A}{R_T^{B(1-\sigma_1)}} \\
&= \frac{\eta}{1-\sigma_1}\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_1} + (1-\eta)\mathcal{S}(1-\sigma_2)\varrho_T\left(a_0\frac{R_T^A}{R_T^B}\right)^{1-\sigma_2} - \xi\tilde{\varphi}_T\frac{R_T^A}{R_T^B} \\
&\equiv \tilde{u}\left(R_T^{AB}\right) - \xi\tilde{\varphi}_T R_T^{AB},
\end{aligned} \tag{6.61}$$

where last equality defines $\tilde{u}(\cdot)$, $R_T^{AB} \equiv R_T^A/R_T^B$, and I define:

$$\tilde{\varphi}_T \equiv \varphi_T\left(R_T^B\right)^{\sigma_1}. \tag{6.62}$$

This change of variables shows that I can equivalently optimize over R_T^{AB}. One complication is that the objective function may not be globally concave in R_T^{AB} if $\sigma_2 < 0$ or if $\varrho'' > 0$. Hence, standard first-order conditions are not sufficient. The standard approach is to construct the concavification of $\tilde{u}\left(R_T^{AB}\right)$ (Carpenter (2000), Cuoco and Kaniel (2006), and Basak, Pavlova, and Shapiro (2007b)), which is the smallest concave function that dominates $\tilde{u}\left(R_T^{AB}\right)$. I call the concavified function $\hat{u}\left(R_T^{AB}\right)$. Details on the construction of this function are provided in Appendix 6.E.2. The resulting optimization problem is given by:

$$\max_{R_T^{AB}} \hat{u}\left(R_T^{AB}\right) - \xi\tilde{\varphi}_T R_T^{AB}. \tag{6.63}$$

Denote the optimal relative terminal asset level by:

$$R_T^{AB\star} = f\left(\xi\tilde{\varphi}_T\right), \tag{6.64}$$

which is decreasing in φ_T as can be shown using the techniques in Basak, Pavlova, and Shapiro (2007b).

Equipped with the optimal terminal relative return, I can compute the time-t return on assets:

$$\begin{aligned} R_t^{A\star} &= E_t\left[\frac{\varphi_T}{\varphi_t}R_T^B f\left(\xi\tilde{\varphi}_T\right)\right] \\ &= R_t^B E_t\left[\frac{\varphi_T}{\varphi_t}\frac{R_T^B}{R_t^B}f\left(\xi\tilde{\varphi}_T\right)\right] \\ &= R_t^B E_t\left[\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t}\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1} f\left(\xi\tilde{\varphi}_T\right)\right] \\ &= R_t^B E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right] E_t\left[\frac{\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}}{E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right]}\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t}f\left(\xi\tilde{\varphi}_T\right)\right] \\ &= R_t^B E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right] E_t^{\mathbb{G}}\left[\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t}f\left(\xi\tilde{\varphi}_T\right)\right], \end{aligned} \tag{6.65}$$

in which I change the measure to an equivalent measure $\mathbb{G}$ via the Radon-Nikodym derivative:

$$\frac{\mathrm{d}\mathbb{G}}{\mathrm{d}\mathbb{P}} \equiv \frac{\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}}{E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right]}. \tag{6.66}$$

All expectations under the equivalent measure $\mathbb{G}$ are denoted by $E_t^{\mathbb{G}}[\cdot]$, while $\mathbb{P}-$expectations are denoted by $E_t[\cdot]$. Note that the first expectation in (6.65) is a deterministic function of the remaining investment horizon, $T-t$. Due to the Markovianity of $(\varphi_t)_{t\geq 0}$, it holds:

$$R_t^{A\star} = R_t^B g\left(\xi,\tilde{\varphi}_t\right), \tag{6.67}$$

with:

$$g\left(\xi,\tilde{\varphi}_t\right) \equiv E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right] E_t^{\mathbb{G}}\left[\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t}f\left(\xi\tilde{\varphi}_T\right)\right]. \tag{6.68}$$

The last step is to compute the optimal investment strategy, $x_t^\star(\varphi_t)$, which is the replicating portfolio of $R_t^{A\star}$. To this end, I match the diffusion term of R_t^A, which is given by $R_t^A x'\Sigma\mathrm{d}Z_t$, with the one of $R_t^{A\star}$. The latter diffusion term takes the form (Ito's lemma):

$$\left(\frac{\partial g\left(\xi,\tilde{\varphi}_t\right)}{\partial\tilde{\varphi}_t}\right) R_t^B\tilde{\varphi}_t\left[-\Lambda' + \sigma_1 e_1'\Sigma\right]\mathrm{d}Z_t + R_t^{A\star}e_1'\Sigma\mathrm{d}Z_t, \tag{6.69}$$

which leads to:

$$x_t^\star(\tilde{\varphi}_t) = \begin{pmatrix} x_t^{B\star}(\tilde{\varphi}_t) \\ x_t^{A\star}(\tilde{\varphi}_t) \end{pmatrix} = e_1 - \sigma_1\left(\frac{\partial g\left(\xi,\tilde{\varphi}_t\right)}{\partial\tilde{\varphi}_t}\right)\frac{\tilde{\varphi}_t}{R_t^{AB\star}}\left[\frac{1}{\sigma_1}\Sigma^{-1}\Lambda - e_1\right], \tag{6.70}$$

with e_1 denoting the first unit vector. Since $\hat{u}\left(R_T^{AB}\right)$ is increasing and concave in R_T^{AB}, it holds that $g\left(\xi,\tilde{\varphi}_T\right)$ is monotonically decreasing in $\tilde{\varphi_T}$. As a result:

$$-\sigma_1\left(\frac{\partial g\left(\xi,\tilde{\varphi}_t\right)}{\partial\tilde{\varphi}_t}\right)\frac{\tilde{\varphi}_t}{R_t^{AB\star}}, \tag{6.71}$$

is positive. If $\sigma_1 > \lambda_B/\sigma_B$, then $x_t^{B\star}(\tilde{\varphi}_t) < 1$, while, by contrast, $x_t^{B\star}(\tilde{\varphi}_t) > 1$ if $\sigma_1 < \lambda_B/\sigma_B$. This implies that the manager can increase of decrease the benchmark weight if she wants to deviate from the herd. It depends on σ_1 how the manager deviates and implies that σ_1 controls passive risk-taking. Quantitatively, the deviation also depends on σ_2, which affects $g(\xi,\tilde{\varphi}_t)$. Lower values of σ_2 will make the utility index more convex, thereby enlarging the risk-shifting region and increasing the manager's tilt away from the benchmark.

The two special cases in which the preferences reduce to the preference specifications in Section 6.4 can be identified easily. If $\eta = 0$, it holds:

$$-\sigma_1 \left(\frac{\partial g\left(\xi, \tilde{\varphi}_t\right)}{\partial \tilde{\varphi}_t} \right) \frac{\tilde{\varphi}_t}{R_t^{AB\star}} = 1, \tag{6.72}$$

and the optimal portfolio simplifies to:

$$x_t^\star = \frac{1}{\sigma_1} \Sigma^{-1} \Lambda. \tag{6.73}$$

When $\eta = 1$, the asset distribution is uniform, and $\sigma_1 = 1$, it holds:

$$-\sigma_1 \left(\frac{\partial g\left(\xi, \tilde{\varphi}_t\right)}{\partial \tilde{\varphi}_t} \right) \frac{\tilde{\varphi}_t}{R_t^{AB\star}} = \frac{1}{\sigma_2}, \tag{6.74}$$

and the optimal portfolio simplifies to:

$$x_t^\star = \frac{1}{\sigma_2} \Sigma^{-1} \Lambda + \left(1 - \frac{1}{\sigma_2}\right) e_1, \tag{6.75}$$

which is the optimal strategy derived in Binsbergen, Brandt, and Koijen (2007).

6.E Econometric approach

This appendix details the construction of the likelihood of mutual funds returns conditional on passive returns. Section 6.E.1 provides the results for the two benchmark models in Section 6.4. Section 6.E.2 constructs the likelihood for the model in Section 6.7 in which the manager care about their relative position in the asset distribution.

6.E.1 Two benchmark models

In both benchmark models, the optimal strategy is a constant-proportions strategy ((6.8) and (6.10)). This implies that the likelihood can be constructed as in Appendix 6.A with:

$$\begin{aligned} \alpha &= x^A \sigma_A \lambda_A, & (6.76)\\ \beta &= x^B, & (6.77)\\ \sigma_\varepsilon &= x^A \sigma_A, & (6.78) \end{aligned}$$

and Θ_C is replaced by $\Theta_A = \{\lambda_A, \gamma\}$.

6.E.2 Status model

Section 6.5 and, in more detail, Koijen (2007) uses the mapping from assets under management to the state-price density that is implied by the martingale method to construct the likelihood of fund returns. This mapping is straightforward to construct in case of a globally concave utility index (Koijen (2007)). This appendix provides the procedure for the status model in which the utility index can feature local convexities. I combine martingale techniques in the presence of local convexities (Basak, Pavlova, and Shapiro (2007b)) with the simplifications in Section 6.D.

I outline the main procedure if there is at most one convex region. The method directly extends to multiple convex regions.

1. Check whether the Lagrangian is globally concave:

$$\tilde{u}\left(R_T^{AB}\right) - \xi \tilde{\varphi}_T R_T^{AB} = \frac{\eta}{1-\sigma_1} \left(a_0 R_T^{AB}\right)^{1-\sigma_1} + (1-\eta)\, \mathcal{S}\left(1-\sigma_2\right) \varrho_T \left(a_0 R_T^{AB}\right)^{1-\sigma_2} - \xi \tilde{\varphi}_T R_T^{AB}, \tag{6.79}$$

for instance, by computing the maximum of $\tilde{u}''\left(R_T^{AB}\right)$. If the maximum is positive, the function features local convexities; otherwise, the Lagrangian is globally concave.

2. If the objective function is not globally concave, I construct its concavification. The concavified function is the smallest function that dominates $\tilde{u}(\cdot)$, see Carpenter (2000), Cuoco and Kaniel (2006), and Basak, Pavlova, and Shapiro (2007b). This means that the convex region is replaced by a chord between R_1 and R_2, where R_1 and R_2 solve ($R_1 < R_2$):

$$\begin{aligned} \tilde{u}\left(R_1\right) &= A + BR_1, && (6.80)\\ \tilde{u}'\left(R_1\right) &= B, && (6.81)\\ \tilde{u}\left(R_2\right) &= A + BR_2 && (6.82)\\ \tilde{u}'\left(R_2\right) &= B. && (6.83) \end{aligned}$$

This results in a system of four equations in four unknowns, which simplifies to:

$$\begin{aligned} \tilde{u}'\left(R_1\right) &= \tilde{u}'\left(R_2\right), && (6.84)\\ \tilde{u}'\left(R_1\right) &= \frac{\tilde{u}\left(R_2\right) - \tilde{u}\left(R_1\right)}{R_2 - R_1}, && (6.85) \end{aligned}$$

which is a system of two equations in two unknowns (R_1 and R_2) only. The concavified function is then defined as:

$$\begin{aligned} \hat{u}\left(R_T^{AB}\right) &= \tilde{u}\left(R_T^{AB}\right) \qquad \text{, if} \quad R_T^{AB} \notin [R_1, R_2]\\ \hat{u}\left(R_T^{AB}\right) &= A + BR_T^{AB} \quad \text{, if} \quad R_T^{AB} \in [R_1, R_2] \end{aligned} \qquad (6.86)$$

This function is, by construction, concave and continuously differentiable. In summary, I use $\tilde{u}(\cdot)$ if the utility index is globally concave and $\hat{u}(\cdot)$ in case of local convexities. I will use $\hat{u}(\cdot)$ in the remainder of the procedure, which can be replaced by $\tilde{u}(\cdot)$ if the utility index is globally concave.

3. Compute the Lagrange parameter, ξ. I need to find ξ that satisfies the budget constraint: $E_0\left[R_T^A \varphi_T\right] = 1$. Appendix 6.D, Equation (6.65) shows that the budget constraint can be written as:

$$1 = E_0\left[\left(R_T^B\right)^{1-\sigma_1}\right] E_0^{\mathbb{G}}\left[\tilde{\varphi}_T f\left(\xi\tilde{\varphi}_T\right)\right], \qquad (6.87)$$

with $f(\xi\tilde{\varphi}_T)$ the optimal terminal asset level relative to the benchmark ($R_T^{AB\star}$) that solves:

$$\max_{R_T^{AB}} \hat{u}\left(R_T^{AB}\right) - \xi\tilde{\varphi}_T R_T^{AB}. \qquad (6.88)$$

The scaled state-price density $\tilde{\varphi}_t$ is defined in (6.62). The first expectation in (6.87) can be computed analytically given the log-normal structure of the financial market:

$$E_0\left[\left(R_T^B\right)^{1-\sigma_1}\right] = \exp\left(\left(1-\sigma_1\right)\left(r + \sigma_P\lambda_P - \frac{1}{2}\sigma_1\sigma_P^2\right)T\right). \qquad (6.89)$$

To compute the second expectation, I use that under $\mathbb{G}$ (by Girsanov's theorem) it holds that (using the Radon-Nikodym derivative in (6.66)):

$$Z_t^{P,\mathbb{G}} \equiv Z_t^P - (1-\sigma_1)\sigma_P t, \qquad (6.90)$$

is a $\mathbb{G}$-Brownian motion. Z_t^A is a $\mathbb{P}$- and $\mathbb{G}$-Brownian motion. This implies that $\tilde{\varphi}_T$ given $\tilde{\varphi}_t$ can be written as ($\tau \equiv T - t$):

$$\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t} = \exp\left(\begin{array}{l} \left[\left(\sigma_1 - 1\right)r - \frac{1}{2}\left(\Lambda'\Lambda + \sigma_1\sigma_P^2\right) + \sigma_1\sigma_P\lambda_P - \left(1-\sigma_1\right)\left(\lambda_P\sigma_P - \sigma_1\sigma_P^2\right)\right]\tau \\ -\left(\lambda_P - \sigma_1\sigma_P\right)Z_{t:T}^{P,\mathbb{G}} \end{array}\right),$$

with $\Delta Z_{t:T} \equiv Z_T - Z_t$. It therefore holds:

$$\log \tilde{\varphi}_T - \log \tilde{\varphi}_t \sim^{\mathbb{G}} N\left(\mu_{\tilde{\varphi},T}, \sigma^2_{\tilde{\varphi},T}\right), \tag{6.91}$$

with:

$$\mu_{\tilde{\varphi},T} \equiv \left[(\sigma_1 - 1)\, r - \frac{1}{2}\left(\Lambda'\Lambda + \sigma_1\sigma_P^2\right) + \sigma_1\sigma_P\lambda_P - (1-\sigma_1)\left(\lambda_P\sigma_P - \sigma_1\sigma_P^2\right)\right]\tau, \tag{6.92}$$

$$\sigma^2_{\tilde{\varphi},T} \equiv \left(\lambda_P - \sigma_1\sigma_P\right)^2\tau. \tag{6.93}$$

The expectation in (6.89) is computed using univariate Gaussian quadrature with six points (see Tauchen and Hussey (1991)). This holds true regardless of the number of Brownian motions driving the uncertainty in the financial market.

4. Solve for the transformed state-price density at each point, $\tilde{\varphi}_t$, in time to match the observed mutual fund return:

$$R_t^{A\star} = R_t^B E_t\left[\left(\frac{R_T^B}{R_t^B}\right)^{1-\sigma_1}\right] E_t^{\mathbb{G}}\left[\frac{\tilde{\varphi}_T}{\tilde{\varphi}_t} f\left(\xi, \tilde{\varphi}_T\right)\right], \tag{6.94}$$

in which the expectations are computed as in the previous step. This completes the mapping from fund returns, $R^{A,T}$, to a time-series of the (transformed) state-price density, $\tilde{\varphi}^T$. Then the log-likelihood contribution follows as a standard application of the change-of-variables theorem:

$$\begin{aligned} \ell\left(R_t^A \mid R^{B,t}, R^{A,t-h}\right) &= \ell\left(R_t^A \mid R_t^B, \log\tilde{\varphi}_{t-h}\right) \\ &= \ell\left(\log\tilde{\varphi}_t \mid R_t^B, \log\tilde{\varphi}_{t-h}\right) + \log\left|\left(\frac{\partial R_{At}^\star}{\partial \log\tilde{\varphi}_t}\right)^{-1}\right|. \end{aligned} \tag{6.95}$$

Note that I do not need to compute the portfolio weights explicitly, as I can compute the likelihood of assets under management directly.

Koijen (2007) provides further details on the exact implementation and explains the method in the model of Section 6.4.2.

6.F Hypothesis testing

I formally test competing models of delegated portfolio management to study which model describes the returns produced by fund managers best. I the testing procedure, I distinguish between nested and non-nested models.

Nested models If the models are nested, the likelihood-ratio test can be used to discriminate between models. Denote by $\mathcal{L}^1$ the log-likelihood corresponding the unconstrained model evaluated at the maximum-likelihood estimates, and $\mathcal{L}^0$ the log-likelihood of the constrained model. The likelihood-ratio statistic:

$$LR = 2\left(\mathcal{L}^1 - \mathcal{L}^0\right), \tag{6.96}$$

follows under the null a chi-squared distribution with the degrees of freedom equal to the number of parameter constraints.

Non-nested models If the models are non-nested, the standard likelihood-ratio test cannot be applied. However, Vuong (1989) develops an alternative test that also uses the likelihood ratio as the main input, and which can be used to test non-nested models.[67] The different dynamic model of delegated management are not necessarily nested. The log-likelihood of fund returns conditional on the benchmark returns

[67] This test is also used in St-Amour (2006) to discriminate between structural consumption-based asset pricing models.

corresponding to Model 1 is denoted by $\mathcal{L}^{(1)}\left(r_A^T \mid r_B^T; \Theta_A\right) = \sum_{t=h}^{T/h} \ell^{(1)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \Theta_A\right)$ and for Model 2 by $\mathcal{L}^{(2)}\left(r_A^T \mid r_B^T; \Theta_A\right) = \sum_{t=h}^{T/h} \ell^{(2)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \Theta_A\right)$. The null hypothesis reads:

$$H_0 : E_0\left[\mathcal{L}^{(1)}\left(r_A^T \mid r_B^T; \Theta_A^{(1),\star}\right)\right] = E_0\left[\mathcal{L}^{(2)}\left(r_A^T \mid r_B^T; \Theta_A^{(2),\star}\right)\right], \tag{6.97}$$

in which $E_0[\cdot]$ denotes the expectation under the true model, and $\Theta_A^{(j),\star}$ the pseudo-true parameters of Model j. The null hypothesis does not require either of the models to be correctly specified. Then under H_0:

$$\frac{\mathcal{L}^{(1)}\left(r_A^T \mid r_B^T; \hat{\Theta}_A^{(1)}\right) - \mathcal{L}^{(2)}\left(r_A^T \mid r_B^T; \hat{\Theta}_A^{(2)}\right) - (p-q)}{\sqrt{\hat{\omega}_T T/h}} \xrightarrow{d,\, H_0} N(0,1), \tag{6.98}$$

with p the number of parameters estimated in Model 1, q the number of parameters for Model 2, and $\hat{\Theta}_A^{(j)}$ the maximum-likelihood estimates of Model j. In addition, $\hat{\omega}_T$ is defined as:

$$\begin{aligned}\hat{\omega}_T \equiv\ & \frac{1}{T/h}\sum_{t=h}^{T/h}\left[\ell^{(1)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \Theta_A^{(1)}\right) - \ell^{(2)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \hat{\Theta}_A^{(2)}\right)\right]^2 \\ & - \left[\frac{1}{T/h}\sum_{t=h}^{T/h}\left[\ell^{(1)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \Theta_A^{(1)}\right) - \ell^{(2)}\left(r_{At} \mid r_B^t, r_A^{t-h}; \hat{\Theta}_A^{(2)}\right)\right]\right]^2.\end{aligned} \tag{6.99}$$

Implementation All tests will be performed at the manager level. This implies that the test will have relatively little power at a per-manager basis. However, I can take advantage of the large cross-section of managers available. If the significance level is set at 5%, I expect to reject the null only for 5% of the managers. If the rejection rate is considerably higher, this provides strong evidence that one of the models provides a better description of mutual fund behavior.

6.G Utility cost calculation

To compute the loss in certainty-equivalent wealth, I need to determine π_1, π_2, and $CEQ(\pi_1, \pi_2)$. First, to compute π_2, I only need to integrate out financial uncertainty. This requires me to obtain the joint distribution of the benchmark, S_T^B, and the state-price density φ_T (see Appendix 6.E.2) to compute W_T. Note that in the derivation of the optimal strategy in Appendix 6.E.2, I require the transformed state-price density $S_t^{B-\gamma}\varphi_t$. As such, it is sufficient to simulate S_T^B and φ_T to compute the payoffs of the active and passive asset, and the investor's wealth in turn. I use Monte Carlo simulations to integrate out the financial uncertainty and I sample the manager's strategy at a monthly frequency. To compute π_1, I also need to integrate out heterogeneity in ability and preferences. I use the empirical distribution function that I estimate in Section 6.8 and the integration is replaced by summation as a result. The expectation with respect to financial risks is again approximated by Monte Carlo simulations. The value function follow directly and hence $CEQ(\pi_1, \pi_2)$. Note that the optimal strategy depends on a_0. For each manager, I take the average over the sample period available for the particular manager and use this value in the simulation exercise.

If the manager uses performance regressions, the computations can be simplified in this case as the value function conditional on the parameters can be computed analytically and is given by ($W_0 = 1$):

$$J(\pi, T, \Theta_A) = \frac{1}{1-\gamma_I}\exp\left((1-\gamma_I)\left(x_I'\bar{\Sigma}\Lambda + r\right) - \frac{\gamma_I(1-\gamma_I)}{2}x_I'\bar{\Sigma}\bar{\Sigma}'x_I\right),$$

with $\bar{\Sigma}$ the volatility matrix of the passive asset and the managed portfolio.

6.H Tables and figures

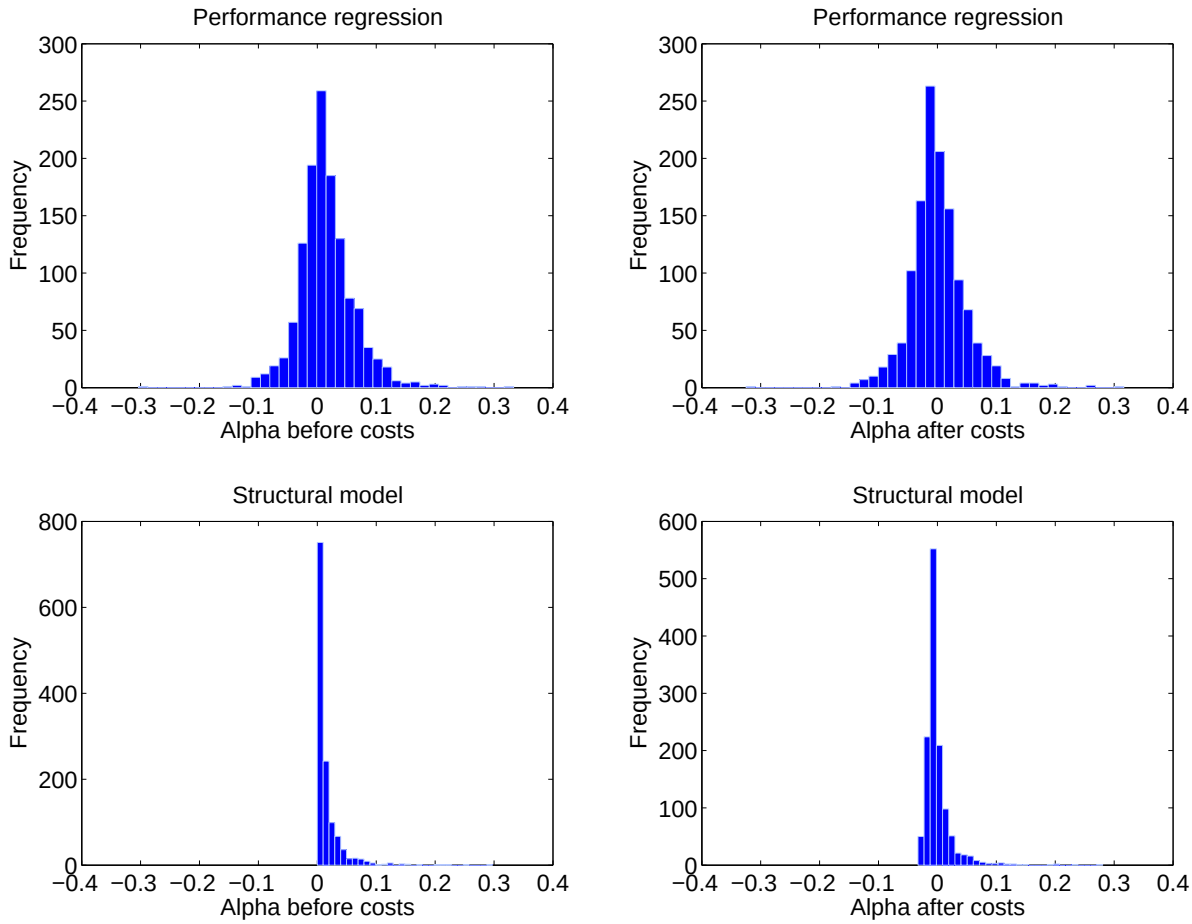


Figure 6.2: Estimated distribution of fund alphas
The figure displays the distribution of fund alphas following from performance regressions (top panels) and from the structural status model (bottom panels) in Section 6.7. The left panels provide the results before fees and expenses, whereas the right panels correspond to the results after fees and expenses. The model is estimated for 1,273 manager-benchmark combinations. The alpha in case of performance regressions is computed as explained in Appendix 6.A.

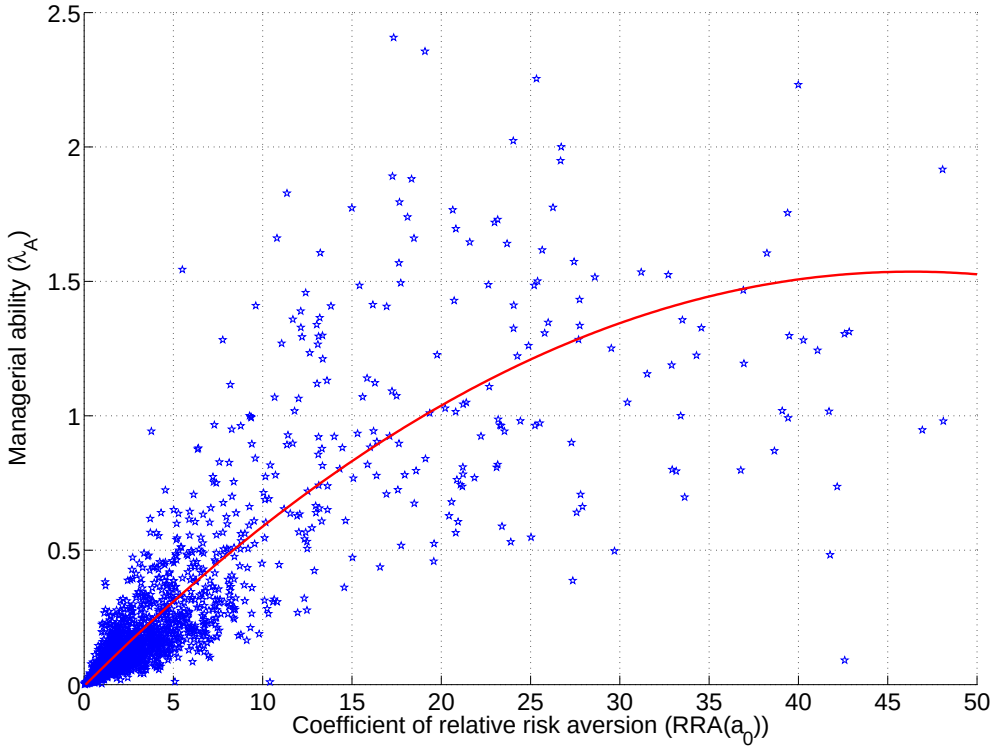


Figure 6.3: Managerial ability and risk aversion
The figure displays the cross-sectional distribution of managerial ability and the coefficient of relative risk aversion that follows from the model in Section 6.7. The model is estimated for 1,273 manager-benchmark combinations. The red line corresponds to a second-order polynomial fitted through the cloud of points to illustrate the relation between managerial ability and the coefficient of relative risk aversion.

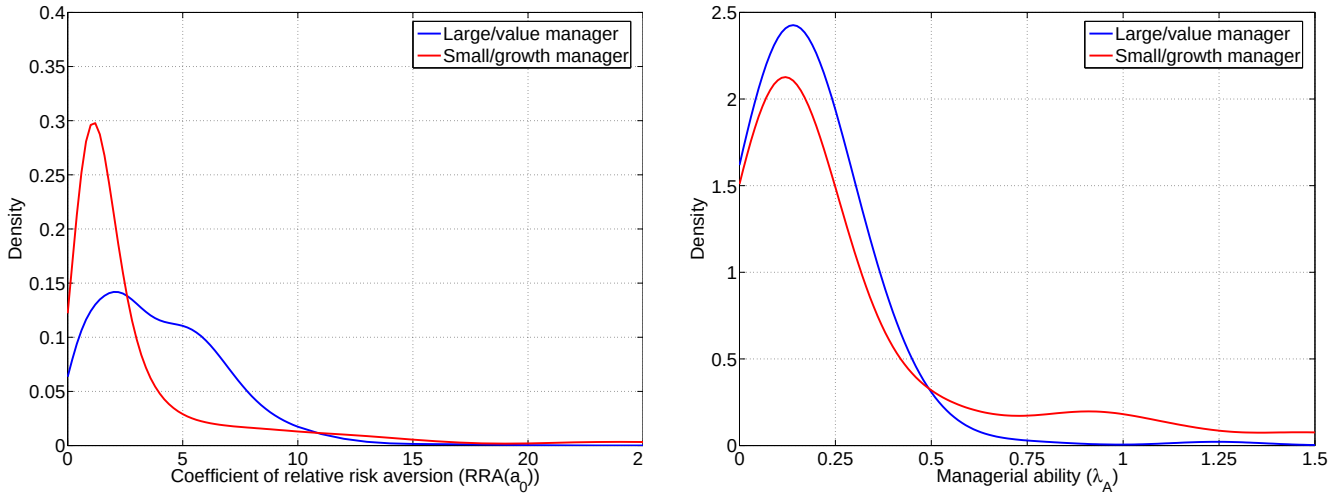


Figure 6.4: Managerial ability and risk aversion across styles
The left panel of this figure displays the distribution of the coefficient of relative risk aversion. The right panel depicts the distribution of managerial ability. The red (blue) lines correspond to the small/growth (large/value) style. The densities are estimated using a standard kernel density estimators based on a normal kernel function.

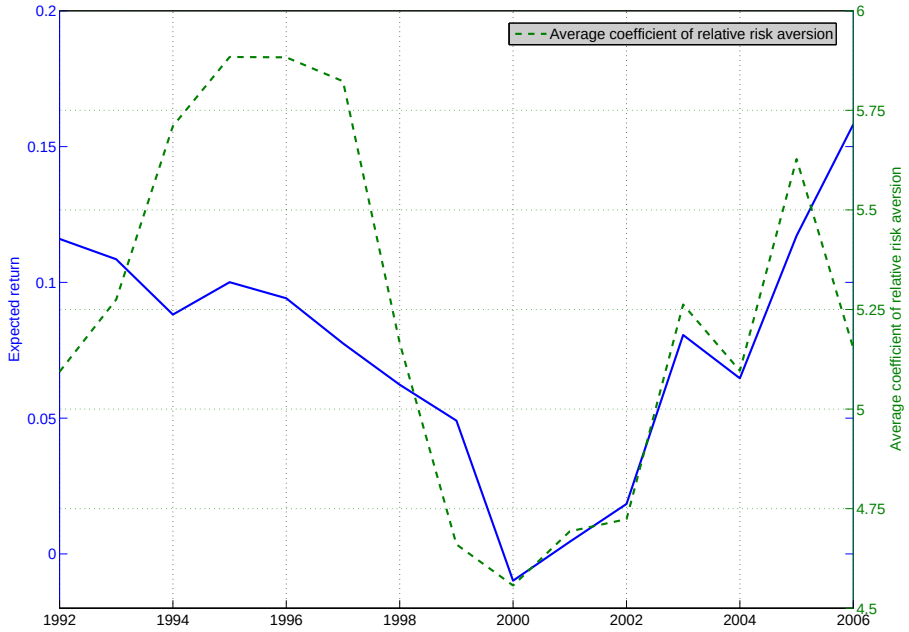


Figure 6.5: Time series of risk aversion and expected returns
The blue line depicts the average time-series variation in the coefficient of relative risk aversion that follows from the status model in Section 6.7. This time series is computed by averaging the coefficients of relative risk aversion in the cross-section of fund managers in each year. The dashed green line corresponds to the time series of the equity risk premium, which is taken from Binsbergen and Koijen (2007).

Mutual fund style	Selected benchmark	Fraction of observations (%)	Number of managers	Fraction of observations (%)	Number of managers
		Years of data ≥ 1		Years of data ≥ 3	
Large-cap/blend	S&P 500	20.1	714	20.3	258
Large-cap/value	Russell 1000 Value	11.7	427	11.7	149
Large-cap/growth	Russell 1000 Growth	11.6	448	11.1	141
Mid-cap/blend	Russell Mid-cap	10.2	383	9.9	126
Mid-cap/value	Russell Mid-cap Value	6.3	228	6.4	82
Mid-cap/growth	Russell Mid-cap Growth	13.7	526	13.5	172
Small-cap/blend	Russell 2000	7.8	291	8.6	110
Small-cap/value	Russell 2000 Value	6.2	200	6.3	80
Small-cap/growth	Russell 2000 Growth	12.4	477	12.2	155
Total		100.0	3,694	100.0	1,273

Table 6.3: **Number of manager-fund combinations per investment style**
The table summarize the number and fraction of manager-fund combinations per investment style. Managers are allocated to a benchmark by performing nine regressions of fund returns in excess of the short rate on excess benchmark returns. I select the benchmark that maximizes the R-squared. The left panel displays the allocation of manager-fund combinations for the full sample, the right panel for manager-fund combinations for which at least 3 years of data is available.

	Mean	St.dev.	Percentiles				
			10%	25%	50%	75%	90%
TNA ($mln)	1,042	3,757	22	54	174	630	2,001
Family TNA ($mln)	14,621	45,837	74	342	1,807	9,050	23,088
Family size	8.3	9.0	1.0	2.0	5.0	11.0	18.0
Expense ratio (%)	1.3	0.5	0.8	1.0	1.3	1.6	1.9
12B-1 fee (bp)	21.5	28.0	0.0	0.0	6.1	34.0	65.0
Total load (%)	2.1	2.4	0.0	0.0	1.0	4.7	5.5
Cash holdings (%)	4.2	4.5	0.0	1.0	3.1	6.3	10.2
Stock holdings (%)	94.9	5.0	87.8	92.4	96.2	98.4	99.7
Manager tenure (years)	5.2	4.5	1.2	2.1	3.9	6.8	10.6
Fund age (years)	11.1	12.5	2.2	3.9	6.9	12.3	27.1
Turnover (%)	89	96	18	36	67	113	175

Table 6.4: **Summary statistics**
The table provides the summary statistics of manager and fund characteristics. I provide summary statistics for the total net assets under management (TNA), total net assets of the fund family (as defined by Chen, Hong, Huang, and Kubik (2004)), family size (the number of funds that belong to the fund family), expense ratio, 12B-1 fees, the total load (the sum of maximum front-end load fees and maximum deferred and rear-end load fees), cash holdings as reported by the fund, stock holdings as reported by the fund (the sum of common and preferred stock), manager's tenure, fund age, and turnover.

Panel A: Structural parameters

Model parameters		λ_A					γ				
λ_A	γ	Mean	Std	10%	50%	90%	Mean	Std	10%	50%	90%
	2	0.10	0.01	0.08	0.10	0.11	2.00	0.09	1.90	2.00	2.11
0.1	5	0.10	0.01	0.08	0.10	0.11	5.01	0.21	4.74	5.00	5.29
	10	0.10	0.01	0.08	0.10	0.11	10.01	0.43	9.48	10.00	10.57
	2	0.20	0.03	0.16	0.19	0.24	2.01	0.18	1.80	2.00	2.24
0.2	5	0.20	0.03	0.16	0.19	0.24	5.03	0.44	4.51	5.00	5.61
	10	0.20	0.03	0.16	0.19	0.24	10.06	0.87	9.01	9.99	11.22
	2	0.30	0.05	0.24	0.29	0.37	2.03	0.27	1.72	2.00	2.39
0.3	5	0.30	0.05	0.24	0.29	0.37	5.08	0.68	4.29	4.99	5.97
	10	0.30	0.05	0.24	0.29	0.37	10.16	1.37	8.58	9.99	11.94

Panel B: Structural and reduced-form estimation of fund alphas

Model parameters			α_{ML}					α_{OLS}				
λ_A	γ	α	Mean	Std	10%	50%	90%	Mean	Std	10%	50%	90%
	2	0.5%	0.49%	0.12%	0.34%	0.47%	0.65%	0.57%	2.95%	-3.18%	0.62%	4.44%
0.1	5	0.2%	0.20%	0.05%	0.14%	0.19%	0.26%	0.23%	1.18%	-1.27%	0.25%	1.77%
	10	0.1%	0.10%	0.02%	0.07%	0.09%	0.13%	0.13%	0.64%	-0.64%	0.12%	0.90%
	2	2.0%	1.96%	0.50%	1.35%	1.90%	2.63%	2.15%	5.90%	-5.34%	2.25%	9.85%
0.2	5	0.8%	0.78%	0.20%	0.54%	0.76%	1.05%	0.86%	2.37%	-2.15%	0.89%	3.93%
	10	0.4%	0.39%	0.10%	0.27%	0.38%	0.53%	0.47%	1.25%	-1.08%	0.45%	2.02%
	2	4.5%	4.46%	1.23%	3.02%	4.30%	6.10%	4.71%	8.84%	-6.54%	4.86%	16.21%
0.3	5	1.8%	1.78%	0.49%	1.21%	1.72%	2.44%	1.89%	3.55%	-2.62%	1.94%	6.52%
	10	0.9%	0.89%	0.25%	0.60%	0.86%	1.22%	0.96%	1.79%	-1.32%	0.96%	3.36%

Table 6.5: **Simulation experiment to compare model-implied and regression-based alphas**
Panel A displays the results of a simulation exercise in which I simulate the model in Section 6.4.1 for three years on a monthly frequency. Managerial ability takes values in $\lambda_A \in \{.1, .2, .3\}$ and the coefficient of relative risk aversion takes values in $\gamma \in \{2, 5, 10\}$. The table provides the mean, standard deviation, and 10%, 50%, and 90% quantiles of the estimates across 2,500 data sets. Both models are estimated by means of likelihood, see Appendix 6.E.1. Panel B displays the fund's alpha that is implied by the structural model, $\alpha_{ML} = \lambda_A^2/\gamma$, or that follows from a standard performance regression, α_{OLS}, see Appendix 6.A.

				Model-implied			Performance regressions		
		γ	λ_A	α	β	σ_ε	α	β	σ_ε
S&P 500	Mean	46.08	1.36	6.27%	1.10	4.48%	0.82%	0.96	4.10%
	Median	31.29	1.37	5.74%	1.10	4.36%	0.67%	0.96	3.75%
	St.dev.	108.15	0.34	3.51%	0.05	2.02%	2.98%	0.11	1.97%
Russell 1000 Value	Mean	38.21	1.62	7.74%	1.15	4.72%	0.30%	0.93	4.09%
	Median	35.94	1.63	7.70%	1.13	4.59%	0.29%	0.92	3.91%
	St.dev.	14.99	0.32	3.24%	0.06	1.65%	2.60%	0.13	1.55%
Russell 1000 Growth	Mean	17.35	0.91	6.00%	1.08	6.19%	1.26%	0.94	5.47%
	Median	15.93	0.89	5.03%	1.06	5.77%	0.99%	0.92	4.99%
	St.dev.	9.30	0.40	4.90%	0.09	3.27%	3.60%	0.18	2.74%
Russell Mid-cap	Mean	23.75	1.48	10.72%	1.19	7.03%	1.11%	0.96	6.47%
	Median	24.17	1.44	9.50%	1.16	6.50%	0.90%	0.95	5.98%
	St.dev.	8.55	0.36	5.89%	0.09	2.90%	4.43%	0.17	2.87%
Russell Mid-cap Value	Mean	27.90	1.75	11.88%	1.23	6.71%	-0.04%	0.94	5.97%
	Median	26.01	1.77	10.81%	1.22	6.62%	0.32%	0.94	6.00%
	St.dev.	8.30	0.29	4.41%	0.07	1.89%	4.15%	0.18	1.77%
Russell Mid-cap Growth	Mean	11.42	0.94	9.27%	1.11	9.30%	0.40%	0.95	7.98%
	Median	10.77	0.93	6.46%	1.08	8.37%	0.59%	0.95	7.76%
	St.dev.	6.01	0.42	7.21%	0.13	4.02%	5.80%	0.22	3.08%
Russell 2000	Mean	19.92	1.35	10.57%	1.11	7.72%	3.55%	0.90	6.78%
	Median	17.80	1.35	9.44%	1.10	7.32%	3.59%	0.93	6.28%
	St.dev.	9.44	0.39	5.62%	0.06	3.13%	5.20%	0.18	2.92%
Russell 2000 Value	Mean	29.01	1.70	11.45%	1.20	6.77%	1.61%	0.95	6.28%
	Median	26.74	1.73	11.67%	1.18	6.53%	1.46%	0.94	5.93%
	St.dev.	12.57	0.29	4.34%	0.11	2.53%	3.68%	0.16	2.49%
Russell 2000 Growth	Mean	7.90	0.64	6.14%	1.05	10.14%	5.49%	0.96	9.10%
	Median	6.86	0.56	4.52%	1.01	9.27%	5.01%	0.97	8.69%
	St.dev.	7.51	0.51	5.82%	0.06	6.31%	6.92%	0.17	4.73%

Table 6.6: **Parameter estimates for the model in Section 6.4.1**
The table summarizes the estimation results for the model in Section 6.4.1. The model is estimated by means of maximum likelihood for 1,273 managers over the period 1992.1 to 2006.12 for all nine investment styles. Fund managers are included when at least three years of return data is available to estimate the models. The first two columns provides the estimates of the structural model, γ and λ_A. Columns three to five provide the implied estimates for the coefficients of a performance regression, α, β, and σ_ε. The last three columns report the results of standard performance regressions (Appendix 6.A). In all cases, I report the cross-sectional mean, median, and standard deviation (St.dev.) of the estimates. The parameters are expressed in annual terms.

				Model-implied			Performance regressions		
		γ	λ_A	α	β	σ_ε	α	β	σ_ε
S&P 500	Mean	4.25	0.18	0.89%	0.95	4.16%	0.82%	0.96	4.10%
	Median	4.19	0.17	0.65%	0.95	3.82%	0.67%	0.96	3.75%
	St.dev.	0.47	0.08	0.87%	0.11	1.99%	2.98%	0.11	1.97%
Russell 1000 Value	Mean	6.41	0.26	1.21%	0.93	4.13%	0.30%	0.93	4.09%
	Median	6.37	0.25	0.99%	0.92	3.90%	0.29%	0.92	3.91%
	St.dev.	0.89	0.10	0.89%	0.13	1.56%	2.60%	0.13	1.55%
Russell 1000 Growth	Mean	2.18	0.12	0.86%	0.94	5.55%	1.26%	0.94	5.47%
	Median	2.05	0.11	0.55%	0.92	5.03%	0.99%	0.92	4.99%
	St.dev.	1.66	0.15	1.54%	0.17	2.77%	3.60%	0.18	2.74%
Russell Mid-cap	Mean	5.20	0.33	2.53%	0.95	6.54%	1.11%	0.96	6.47%
	Median	5.08	0.30	1.80%	0.95	6.04%	0.90%	0.95	5.98%
	St.dev.	0.92	0.15	2.39%	0.17	2.85%	4.43%	0.17	2.87%
Russell Mid-cap Value	Mean	7.81	0.46	2.94%	0.93	6.02%	-0.04%	0.94	5.97%
	Median	7.33	0.42	2.55%	0.93	6.03%	0.32%	0.94	6.00%
	St.dev.	3.13	0.20	1.77%	0.18	1.74%	4.15%	0.18	1.77%
Russell Mid-cap Growth	Mean	2.04	0.16	1.49%	0.95	8.11%	0.40%	0.95	7.98%
	Median	1.94	0.15	1.19%	0.95	7.86%	0.59%	0.95	7.76%
	St.dev.	0.49	0.07	1.16%	0.22	3.15%	5.80%	0.22	3.08%
Russell 2000	Mean	3.24	0.22	1.74%	0.90	6.97%	3.55%	0.90	6.78%
	Median	2.92	0.20	1.35%	0.93	6.38%	3.59%	0.93	6.28%
	St.dev.	1.77	0.10	1.78%	0.17	2.94%	5.20%	0.18	2.92%
Russell 2000 Value	Mean	6.19	0.38	2.70%	0.95	6.32%	1.61%	0.95	6.28%
	Median	6.11	0.37	2.13%	0.94	5.92%	1.46%	0.94	5.93%
	St.dev.	1.13	0.14	1.87%	0.16	2.50%	3.68%	0.16	2.49%
Russell 2000 Growth	Mean	1.16	0.11	1.19%	0.96	9.38%	5.49%	0.96	9.10%
	Median	1.12	0.10	0.87%	0.97	8.89%	5.01%	0.97	8.69%
	St.dev.	0.22	0.05	1.27%	0.17	4.79%	6.92%	0.17	4.73%

Table 6.7: **Parameter estimates for the model in Section 6.4.2**
The table summarizes the estimation results for the model in Section 6.4.2. The model is estimated by means of maximum likelihood for 1,273 managers over the period 1992.1 to 2006.12 for all nine investment styles. Fund managers are included when at least three years of return data is available to estimate the models. The first two columns provides the estimates of the structural model, γ and λ_A. Columns three to five provide the implied estimates for the coefficients of a performance regression, α, β, and σ_ε. The last three columns report the results of standard performance regressions (Appendix 6.A). In all cases, I report the cross-sectional mean, median, and standard deviation (St.dev.) of the estimates. The parameters are expressed in annual terms.

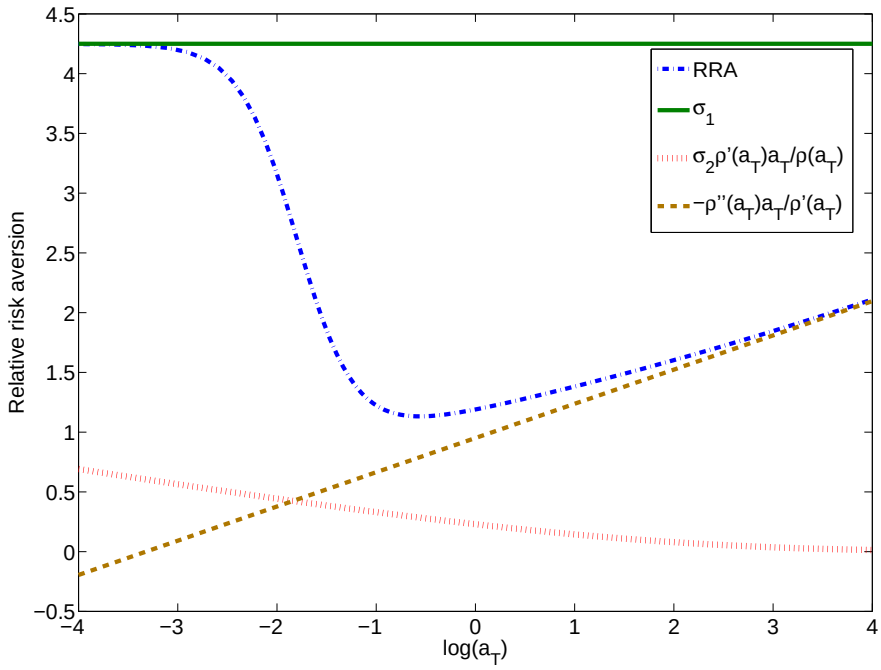


Figure 6.6: Fund size and the coefficient of relative risk aversion ($RRA(a_T)$)
The figure displays the coefficient of relative risk aversion (blue dashed-dotted line) on the vertical axis as a function of the fund's relative size on the horizontal axis for the model in Section 6.7. The coefficient of relative risk aversion is decomposed into three components, see Equation (6.28): (i) σ_1 (green solid line), (ii) $\sigma_2 \varrho'(a_T) a_T / \varrho(a_T)$ (red dotted line), and (iii) $-\varrho''(a_T) a_T / \varrho'(a_T)$ (brown dashed line).

$\sigma_1 = 4.00$, $a_0 = 1$				$\sigma_1 = 3.75$, $\sigma_2 = .5$				$\sigma_1 = 4.25$, $\sigma_2 = .5$			
σ_2	RRA	x^B	x^A	$\rho_0(a_0)$	RRA	x^B	x^A	$\rho_0(a_0)$	RRA	x^B	x^A
-1	0.50	100%	155%	10%	2.57	111%	33%	10%	3.63	92%	23%
0	0.96	100%	78%	20%	1.27	121%	63%	20%	1.66	83%	52%
0.5	1.19	100%	62%	30%	1.10	123%	68%	30%	1.16	78%	66%
1.5	1.64	100%	44%	40%	1.14	121%	64%	40%	1.15	79%	64%
2.5	2.10	100%	34%	50%	1.22	120%	60%	50%	1.22	80%	60%
5	3.25	100%	22%	60%	1.31	119%	56%	60%	1.31	81%	56%
10	5.54	100%	13%	70%	1.41	117%	52%	70%	1.41	83%	52%
20	10.13	100%	7%	80%	1.54	116%	47%	80%	1.54	84%	47%
30	14.72	100%	5%	90%	1.74	114%	42%	90%	1.74	86%	42%

Table 6.8: **Fund status, risk aversion, and risk-taking**
The first four columns display the optimal initial allocation to the benchmark portfolio (x^B) and the active portfolio (x^A) as well as the Arrow-Pratt measure of relative risk aversion for different values of σ_2. The next four columns display the optimal initial strategies and coefficient of relative risk aversion for different initial values of assets under management, a_0, expressed in terms of percentile rank ($\varrho_0(a_0)$) if $\sigma_1 = 3.75$. The last four columns provide the results for $\sigma_1 = 4.25$. I set $\eta = .0005$, $\sigma_2 = .5$ and $\lambda_A = .15$ for the results in the last eight columns. The volatility of the active portfolio, $\sigma_A = 20\%$. The market parameters and the parameters describing the asset distribution are calibrated on the basis of the S&P 500 so that $\lambda_B/\sigma_B = 4$. The short rate is set to $r = 5\%$.

		σ_1	σ_2	λ_A	RRA			σ_1	σ_2	λ_A	RRA
S&P 500	Mean	4.32	11.04	0.23	5.96	R. Mid-cap G.	Mean	2.09	8.34	0.26	3.47
	Median	4.09	0.77	0.11	3.17		Median	1.89	0.77	0.12	1.52
	St.dev.	1.13	28.08	0.33	9.27		St.dev.	0.98	22.08	0.34	5.02
R. 1000 V.	Mean	6.35	11.44	0.21	5.66	R. 2000	Mean	3.61	9.63	0.37	6.04
	Median	6.06	1.56	0.16	3.95		Median	2.88	0.99	0.17	2.28
	St.dev.	1.38	28.12	0.24	7.04		St.dev.	2.16	23.79	0.49	8.69
R. 1000 G.	Mean	2.18	9.23	0.22	4.67	R. 2000 V.	Mean	6.15	7.52	0.26	4.81
	Median	1.96	0.53	0.08	1.66		Median	5.81	1.00	0.15	2.48
	St.dev.	1.66	25.11	0.36	8.83		St.dev.	1.29	21.78	0.28	5.68
R. Mid-cap	Mean	5.29	8.15	0.30	5.01	R. 2000 G.	Mean	1.54	10.60	0.43	5.49
	Median	4.93	1.29	0.18	2.71		Median	1.14	0.99	0.16	1.49
	St.dev.	1.79	22.94	0.36	7.36		St.dev.	1.57	22.13	0.58	8.21
R. Mid-cap V.	Mean	7.51	6.09	0.28	5.06	Overall	Mean	4.05	9.50	0.28	5.16
	Median	7.03	1.49	0.21	4.37		Median	4.04	0.99	0.14	2.51
	St.dev.	1.83	20.31	0.26	5.46		St.dev.	2.41	24.57	0.38	7.69

Table 6.9: **Parameter estimates for the status model in Section 6.7**
The model is estimated for by means of maximum likelihood for 1,273 managers over the period 1992.1 to 2006.12 for all nine investment styles. I also report the results across all styles ("Overall"). Fund managers are included when at least three years of return data is available to estimate the models. The first three columns provides the estimates of the structural model, σ_1, σ_2, and λ_A. The last column reports the implied coefficient of relative risk aversion. λ_A is expressed in annual terms. In all cases, I report the cross-sectional mean, median, and standard deviation (St.dev.) of the estimates. R. abbreviates Russell, V. Value, and G. Growth.

	$\log(\lambda_A)$		$\log(RRA)$	
	Estimate	T-statistic	Estimate	T-statistic
Log(TNA)	-8.87%**	-2.55	-9.99%**	-2.93
Tenure	7.27%**	2.19	4.10%	1.26
Turnover	6.36%**	2.01	0.11%	0.04
Log(Expenses)	5.04%	1.16	-9.07%**	-2.13
Stock holdings	-6.37%**	-2.17	-6.47%**	-2.24
Loads	-3.41%	-1.00	1.17%	0.35
12B-1 fees	0.04%	0.01	4.38%	1.07
Log(Family TNA)	0.10%	0.03	3.30%	1.00
Fund age	3.53%	1.10	2.48%	0.79
R-squared	13.0%		6.6%	

Table 6.10: **Heterogeneity in risk aversion and ability**
The table displays results of multiple cross-sectional regressions of managerial ability and risk aversion on observable manager and fund characteristics. The characteristics include the fund's total net assets, the manger's tenure, turnover, expenses, the investment in common and preferred stocks, loads, 12B-1 fees, the family's total net assets, and the fund's age. The cross-sectional regressions include style dummies. The standard errors used to compute t-statistics are robust to heteroscedasticity. ** indicates statistical significance at the 5% level.

	Relative-return preferences		Preferences for assets under management		Performance regressions	
Investment style	10%	5%	10%	5%	10%	5%
S&P 500	77.1%	68.6%	36.0%	29.8%	20.9%	14.0%
Russell 1000 Value	91.3%	86.6%	43.6%	37.6%	21.5%	12.8%
Russell 1000 Growth	61.7%	53.2%	33.3%	28.4%	17.7%	10.6%
Russell Mid-cap	82.5%	72.2%	34.1%	25.4%	7.9%	4.8%
Russell Mid-cap Value	90.2%	89.0%	35.4%	26.8%	9.8%	4.9%
Russell Mid-cap Growth	58.1%	47.1%	40.7%	29.1%	18.0%	8.7%
Russell 2000	68.2%	52.7%	36.4%	26.4%	10.0%	3.6%
Russell 2000 Value	90.0%	85.0%	31.3%	25.0%	18.8%	15.0%
Russell 2000 Growth	38.7%	31.6%	45.8%	37.4%	16.1%	12.9%
Overall	71.2%	62.9%	37.9%	30.2%	16.6%	10.3%

Table 6.11: **Testing competing models**
The table displays the results testing competing models to describe fund returns. The models under the null are: (i) relative-return preferences (Section 6.4.1), (ii) preferences for assets under management (Section 6.4.2), and (iii) reduced-form performance regressions (Appendix 6.A). The alternative is the status model (Section 6.7). For the models that are nested, I use the likelihood-ratio test. To compare non-nested models, I use the test developed in Vuong (1989), see Appendix 6.F. I test the models at the manager level and report the average number of rejections at either the 5% or 10% significance level. As such, a model is rejected if the average number of rejections exceeds 5% or 10%.

	Fraction significant at the 5% level	
Investment style	Performance regression	Status model
S&P 500	5.4%	7.4%
Russell 1000 Value	4.7%	6.7%
Russell 1000 Growth	6.4%	6.4%
Russell Mid-cap	6.3%	15.9%
Russell Mid-cap Value	2.4%	12.2%
Russell Mid-cap Growth	5.8%	19.2%
Russell 2000	23.6%	21.8%
Russell 2000 Value	2.5%	11.3%
Russell 2000 Growth	27.1%	21.3%

Table 6.12: **Testing for the fraction of skilled managers**
The table reports the fraction of managers that significantly recuperates their fees and expenses at the 5% level. I use either performance regressions or the status model to estimate the fund's alpha. I subsequently test whether the alpha, after fees and expenses, reliably exceeds zero. For the status model, the standard errors are computed using the delta method. The main text provides further details.

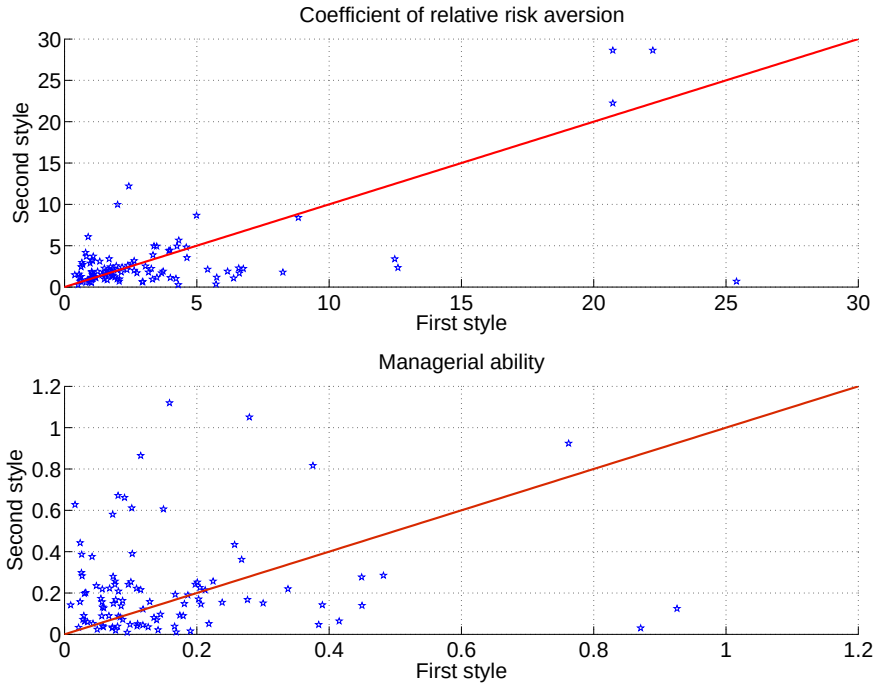


Figure 6.7: Cross-sectional stability of ability and risk aversion
The top panel compares the estimates for the coefficient of relative risk aversion across styles for managers who simultaneously manage funds in different styles. The bottom panel displays the same results for managerial ability. The red line corresponds to the 45-degree line along which the estimates ideally line up. The sample contains 105 managers that are active in multiple styles and that have at least three years of data.

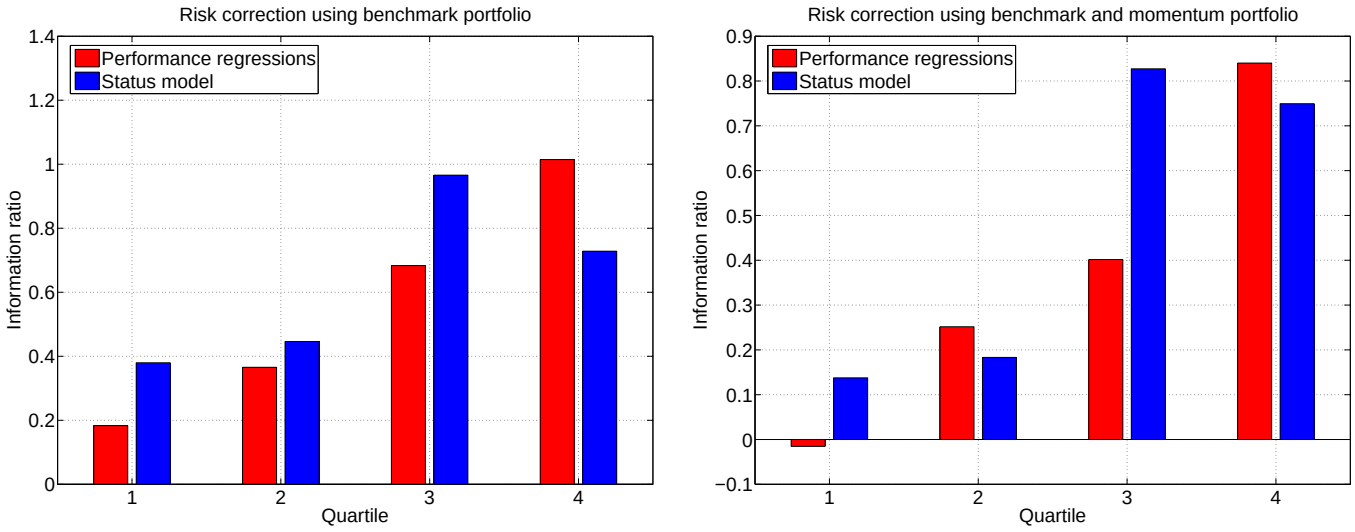


Figure 6.8: Performance persistence
The figure provides the results of a performance persistence test. I use either performance regressions (red bars) or the status model (blue bars) to sort funds into quartiles based on a three-year selection period. I then form equally-weighted portfolios of funds in a particular quartile and compute the portfolio return and the corresponding (equally-weighted) benchmark return. I hold the portfolios for one year. This leads to return series over the full sample period, which I use to compute the (annualized) information ratio (left panel). The information ratio is computed as the intercept of a performance regression, divided by the standard deviation of the residual. To annualize the information ratio, it is multiplied by the square root of twelve. As an alternative, I include the momentum factor to correct for the returns on passive strategies (right panel).

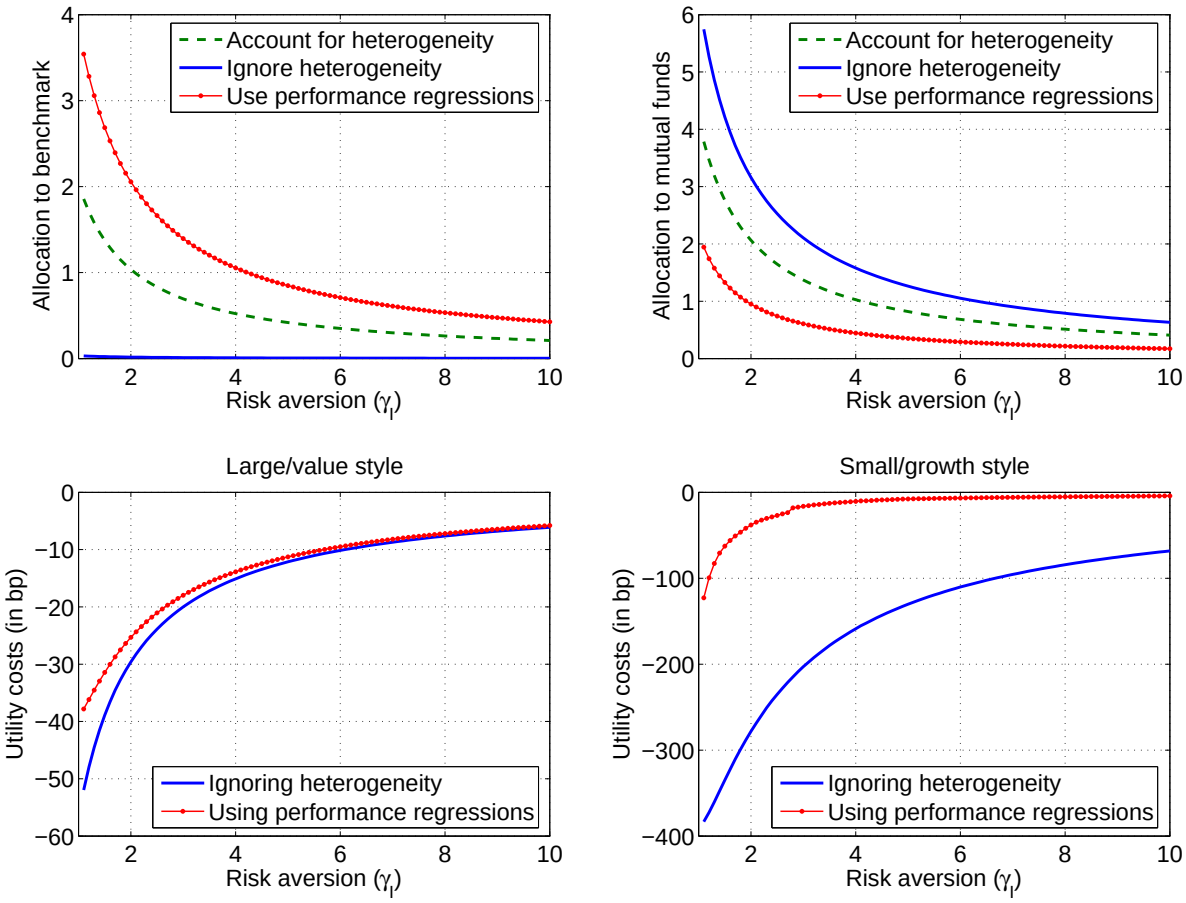


Figure 6.9: Delegated investment management
The top panels display the optimal allocation to the style benchmark (left) and actively-managed mutual funds (right) for an investor in the large/value style. The optimal allocations are determined for three investors and indicated on the vertical axes, the coefficient of relative risk aversion is displayed on the horizontal axis. The green dashed lines in corresponds to the strategy that accounts for heterogeneity using the status model. This case is first-best. The blue solid lines represent the optimal strategy if the investor ignores heterogeneity and uses the average parameter values. The red dotted line corresponds to an investor that (inefficient) performance regressions to account for heterogeneity. The bottom panels compute the loss in certainty-equivalent wealth for both sub-optimal strategies. The solid blue lines corresponds to the investor that ignores heterogeneity, the red dotted line to an investor that uses performance regressions. The bottom left panel corresponds to the large/value investment style; the bottom right panel to small/growth.

References

Acharya, V. V., and J. N. Carpenter (2002): "Corporate Bond Valuation and Hedging with Stochastic Interest Rates and Endogenous Bankruptcy," *The Review of Financial Studies*, 15, 1355–1383.

Admati, A. R., and P. Pfleiderer (1997): "Does It All Add Up? Benchmarks and the Compensation of Active Portfolio Managers," *Journal of Business*, 70, 323–350.

Agarwal, S., J. C. Driscoll, X. Gabaix, and D. Laibson (2007): "The Age of Reason: Financial Decisions Over the Lifecycle," Working Paper.

Ait-Sahalia, Y. (2002): "Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach," *Econometrica*, 70, 223–262.

——— (2007): "Closed-Form Likelihood Expansions for Multivariate Diffusions," *Annals of Statistics*, Forthcoming.

Aït-Sahalia, Y., and M. W. Brandt (2001): "Variable Selection for Portfolio Choice," *Journal of Finance*, 56, 1297–1350.

Amihud, Y., and C. M. Hurvich (2004): "Predictive Regressions: A Reduced-Bias Estimation Method," *Journal of Financial and Quantitative Analysis*, 39, 813–841.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutstrom (2005): "Preference Heterogeneity in Experiments: Comparing the Field and Lab," Working Paper University of Durham.

Ang, A., and G. Bekaert (2005): "The Term Structure of Real Rates and Expected Inflation," Working Paper Columbia University.

——— (2007): "Stock Return Predictability: Is it there?," *Review of Financial Studies*, 20, 651–707.

Ang, A., G. Bekaert, and M. Wei (2006): "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?," *Journal of Monetary Economics*, Forthcoming.

Ang, A., and J. Liu (2004): "How to Discount Cashflows with Time-Varying Expected Returns," *Journal of Finance*, 59, 2745–2783.

——— (2006): "Risk, Return, and Dividends," *Journal of Financial Economics,* forthcoming.

Ang, A., and M. Piazzesi (2003): "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables," *Journal of Monetary Economics*, 50, 745–787.

Assem, J. v. d., G. Baltussen, T. Post, and R. H. Thaler (2007): "Deal or No Deal? Decision Making Under Risk in a Large-Payoff Game Show," *American Economic Review*, Forthcoming.

Avramov, D. (2002): "Stock Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423–458.

——— (2004): "Stock Return Predictability and Asset Pricing Models," *The Review of Financial Studies*, 17, 699–738.

Avramov, D., and T. Chordia (2006): "Predicting Stock Returns," *Journal of Financial Economics*, 82, 387–415.

Avramov, D., and R. Wermers (2006): "Investing in Mutual Funds when Returns are Predictable," *Journal of Financial Economics*, 81, 339–377.

Baker, M., R. Greenwood, and J. Wurgler (2003): "The Maturity of Debt Issues and Predictable Variation in Bond Returns," *Journal of Financial Economics*, 70, 261–291.

Baker, M., R. Taliaferro, and J. Wurgler (2006): "Predicting Returns with Managerial Decision Variables: Is There a Small-sample Bias?," *The Journal of Finance*, 61, 1711–1729.

Baker, M., and J. Wurgler (2000): "The Equity Share in New Issues and Aggregate Stock Returns," *Journal of Finance*, 55, 2219–2258.

Baks, K., A. Metrick, and J. Wachter (2001): "Should Investors Avoid All Actively Managed Mutual Funds? A Study in Bayesian Performance Evaluation," *Journal of Finance*, 56, 45–85.

Baks, K. P. (2006): "On the Performance of Mutual Fund Managers," *Journal of Finance*, Forthcoming.

Bakshi, G., P. Carr, and L. Wu (2007): "Stochastic Risk Premiums, Stochastic Skewness in Currency Options, and Stochastic Discount Factors in International Economies," *Journal of Financial Economics,* forthcoming.

Bakshi, G., and Z. Chen (1996): "The Spirit of Capitalism and Stock-market Prices," *American Economic Review*, 86, 133–157.

Bakshi, G., and N. Ju (2005): "A Refinement to Ait-Sahalia's (2002) "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach"," *Journal of Business*, 78, 2037–2052.

Balduzzi, P., and A. Lynch (1999): "Transaction Costs and Predictability: Some Utility Cost Calculations," *Journal of Financial Economics*, 52, 47–78.

Bansal, R., R. Gallant, and G. Tauchen (2007): "Rational Pessimism, Rational Exuberance, and Asset Pricing Models," *Review of Economic Studies, forthcoming.*

Bansal, R., D. Kiku, and A. Yaron (2006): "Long Run Risks: Estimation and Inference," Unpublished paper, Duke University and the Wharton School.

Bansal, R., and A. Yaron (2004): "Risks for the Long-Run: A Potential Resolution of Asset Pricing Puzzles," *Journal of Finance*, 59(4), 1481–1509.

Barber, B. M., T. Odean, and L. Zheng (2005): "Out of Sight, Out of Mind: The Effects of Expenses on Mutual Fund Flows," *Journal of Business*, 78, 2095–2119.

Barberis, N. (2000): "Investing for the Long Run When Returns Are Predictable," *The Journal of Finance*, 55, 225–264.

Barillas, J., and J. Fernández-Villaverde (2006): "A Generalization of the Endogenous Grid Method," *Journal of Economic Dynamics and Control*, Forthcoming.

Barry, C. B., and L. T. Starks (1984): "Investment Management and Risk Sharing with Multiple Managers," *Journal of Finance*, 39, 477–491.

Basak, S., and D. Makarov (2007): "Strategic Asset Allocation with Relative Performance Concerns," Working paper London Business School.

Basak, S., A. Pavlova, and A. Shapiro (2007a): "Offsetting the Incentives: Benefits of Benchmarking in Money Management," Working Paper London Business School.

——— (2007b): "Optimal Asset Allocation and Risk Shifting in Money Management," *Review of Financial Studies*, Forthcoming.

Basak, S., A. Shapiro, and L. Tepla (2006): "Risk Management with Benchmarking," *Management Science*, 52, 542–557.

Becker, C., W. Ferson, D. H. Myers, and M. J. Schill (1999): "Conditional Market Timing with Benchmark Investors," *Journal of Financial Economics*, 52, 119–148.

Becker, G. S., K. M. Murphy, and I. Werning (2005): "The Equilibrium Distribution of Income and the Market for Status," *Journal of Political Economy*, 113, 282–310.

Bekaert, G., E. Engstrom, and S. R. Grenadier (2001): "Stock and Bond Pricing in an Affine Economy," NBER Working Paper No. 7346.

——— (2005): "Stock and Bond Returns with Moody Investors," Working Paper, Columbia University.

Benzoni, L., P. Collin-Dufresne, and R. S. Goldstein (2006): "Portfolio Choice over the Life-Cycle when the Stock and Labor Markets are Cointegrated," *The Journal of Finance*, Forthcoming.

Bergstrom, A. R. (1984): "Continuous Time Stochastic Models and Issues of Aggregation over Time," in *Handbook of Econometrics*.

Berk, J. B., and R. C. Green (2004): "Mutual Fund Flows and Performance in Rational Markets," *Journal of Political Economy*, 112, 1269–1295.

Berkovec, J. A., D. J. Kogut, and F. E. Nothaft (2001): "Determinants of the ARM Share of FHA and Conventional Lending," *The Journal of Real Esate Finance and Economics*, 22, 23–41.

Binsbergen, J. H., and M. W. Brandt (2007): "Optimal Asset Allocation in Asset Liability Management," Working Paper Duke University.

Binsbergen, J. H. v., M. W. Brandt, and R. S. Koijen (2007): "Optimal Decentralized Investment Management," *Journal of Finance*, Forthcoming.

Binsbergen, J. V., and R. S. Koijen (2007): "Predictive Regressions: A Present-value Approach," Working Paper, Duke University and Tilburg University.

Black, F., and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637–654.

Blake, D. (1999): "Annuity Markets: Problems and Solutions," *The Geneva Papers on Risk and Insurance*, 24, 358–375.

Blake, D., A. J. Cairns, and K. Dowd (2003): "Pensionmetrics 2: Stochastic Pension Plan Design during the Distribution Phase," *Insurance: Mathematics and Economics*, 33, 29–47.

Bliss, R. R. (1997): "Testing Term Structure Estimation Methods," in *Vol. 9 of Advances in Futures and Options Research*, ed. by P. Boyle, G. Pennacchi, and P. Ritchken, pp. 197–231. JAI Press, Greenwich, Connecticut.

Bodie, Z., R. C. Merton, and W. F. Samuelson (1992): "Labor Supply Flexibility and Portfolio Choice in a Life-Cycle Model," *Journal of Economic Dynamics and Control*, 16, 427–449.

Bodie, Z., and J. Pesando (1983): "Retirement Annuity Design in an Inflationary Climate," in *Financial Aspects of the U.S. Pension System*, ed. by Z. Bodie, and J. B. Shoven. Chicago: University of Chicago Press.

Boudoukh, J., R. Michaely, M. Richardson, and M. Roberts (2004): "On the Importance of Measuring Payout Yield: Implications for Empirical Asset Pricing," *NBER Working Paper*, 10651.

Boudoukh, J., M. Richardson, R. Stanton, and R. F. Whitelaw (2004): "The Economics of Assets Management," Working Paper NYU Stern.

Boudoukh, J., R. F. Whitelaw, M. Richardson, and R. Stanton (1997): "Pricing Mortgage-Backed Securities in a Multi-Factor Interest-Rate Environment: A Multivariate Density Approach," *The Review of Financial Studies*, 10, 405–446.

Boulier, J.-F., S. Huang, and G. Taillard (2001): "Optimal Management under Stochastic Interest Rates: The Case of a Protected Defined Contribution Pension Fund," *Insurance: Mathematics and Economics*, 28, 173–189.

Brandt, M. W. (1999): "Estimating Portfolio and Consumption Choice: A Conditional Euler Equations Approach," *The Journal of Finance*, 54, 1609–1645.

Brandt, M. W., A. Goyal, P. Santa-Clara, and J. R. Stroud (2005): "A Simulation Approach to Dynamic Portfolio Choice with an Application to Learning About Return Predictability," *The Review of Financial Studies*, 18, 831–873.

Brandt, M. W., and Q. Kang (2004): "On the Relation Between the Conditional Mean and Volatility of Stock Returns: A Latent VAR Approach," *Journal of Financial Economics*, 72, 217–257.

Brandt, M. W., and P. Santa-Clara (2002): "Simulated Likelihood Estimation of Diffusions with an Application to Exchange Rate Dynamics in Incomplete Markets," *Journal of Financial Economics*, 63, 161–210.

——— (2006): "Dynamic Portfolio Selection by Augmenting the Asset Space," *The Journal of Finance*, 61, 2187–2218.

Brav, A., J. Graham, C. Harvey, and R. Michaely (2005): "Payout Policy in the 21st Century," *Journal of Financial Economics*, 77, 483–527.

Brennan, M. J. (1993): "Agency and Asset Pricing," Working paper, UCLA.

Brennan, M. J., and Y. Xia (2000): "Stochastic Interest Rates and the Bond-Stock Mix," *European Finance Review*, 4, 197–210.

Brennan, M. J., and Y. Xia (2001): "Stock Price Volatility and the Equity Premium," *Journal of Monetary Economics*, 47, 249–283.

Brennan, M. J., and Y. Xia (2002): "Dynamic Asset Allocation under Inflation," *The Journal of Finance*, 57, 1201–1238.

——— (2005): "Persistence, Predictability, and Portfolio Planning," Working Paper, the Anderson School and the University of Pennsylvania.

Brown, J. R. (2001): "Private Pensions, Mortality Risk, and the Decision to Annuitize," *Journal of Public Economics*, 82, 29–62.

Brown, J. R., and J. M. Poterba (2000): "Joint Life Annuities and Annuity Demand by Married Couples," *The Journal of Risk and Insurance*, 67, 527–554.

Brown, K. C., W. Harlow, and L. T. Starks (1996): "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry," *Journal of Finance*, 51, 85–110.

Brown, Jeffrey R., O. S. M., and J. M. Poterba (2001): "The Role of Real Annuities and Indexed Bonds in an Individual Accounts Retirement Program," in *Risk Aspects of Investment-Based Social Security Reform*, ed. by J. Y. Campbell, and M. Feldstein. Chicago: University of Chicago Press.

Brown, S. J., and W. N. Goetzmann (1995): "Performance Persistence," *Journal of Finance*, 50, 679–698.

——— (1997): "Mutual Fund Styles," *Journal of Financial Economics*, 43, 373–399.

Browne, S. (1999): "Beating a Moving Target: Optimal Portfolio Strategies for Outperforming a Stochastic Benchmark," *Finance and Stochastics*, 3, 275–294.

——— (2000): "Risk Constrained Dynamic Active Portfolio Management," *Management Science*, 46, 1188–1199.

Browne, S., M. A. Milevsky, and T. S. Salisbury (2003): "Asset Allocation and the Liquidity Premium of Illiquid Annuities," *The Journal of Risk and Insurance*, 70, 509–526.

Brunnermeier, M., and C. Julliard (2006): "Money Illusion and Housing Frenzies," Working Paper, Princeton University.

Buraschi, A., and A. Jiltsov (2005): "Time-varying inflation risk premia and the expectations hypothesis: A monetary model of the treasury yield curve," *Journal of Financial Economics*, 75, 429–490.

Buraschi, A., P. Porchia, and F. Trojani (2007): "Correlation Risk and Optimal Portfolio Choice," Working Paper, Tanaka Business School.

Burnside, C. (1998): "Solving Asset Pricing Models with Gaussian Shocks," *Journal of Economic Dynamics and Control*, 22, 329–340.

Busse, J. A. (2001): "Another Look at Mutual Fund Tournaments," *Journal of Financial and Quantitative Analysis*, 36, 53–73.

Butler, A. W., G. Grullon, and J. P. Weston (2006): "Can Managers Successfully Time the Maturity Structure of Their Debt Issues?," *The Journal of Finance*, 61, 1731–1758.

Cairns, A. J., D. Blake, and K. Dowd (2006): "Stochastic Lifestyling: Optimal Dynamic Asset Allocation for Defined Contribution Pension Plans," *Journal of Economic Dynamics and Control*, 30, 843–877.

Campbell, J. Y. (1991): "A Variance Decomposition for Stock Returns," *Economic Journal*, 101, 157–179.

——— (2006): "Household Finance," *The Journal of Finance*, 61, 1553–1604.

Campbell, J. Y., and J. Ammer (1993): "What Moves the Stock and Bond Markets? A Variance Decomposition for Long-Term Asset Returns," *Journal of Finance*, 48, 3–37.

Campbell, J. Y., Y. L. Chan, and L. Viceira (2003): "A Multivariate Model of Strategic Asset Allocation," *Journal of Financial Economics*, 67, 41–80.

Campbell, J. Y., and J. Cocco (2003): "Household Risk Management and Optimal Mortgage Choice," *Quarterly Journal of Economics*, 118, 1449–1494.

Campbell, J. Y., and J. H. Cochrane (1999): "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 107(2), 205–251.

Campbell, J. Y., and R. J. Shiller (1988): "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1, 195–227.

——— (1991): "Yield Spreads and Interest Rate Movements: A Bird's Eye View," *Review of Economic Studies*, 58, 495–514.

Campbell, J. Y., and S. Thompson (2007): "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *Review of Financial Studies*, Forthcoming.

Campbell, J. Y., and L. Viceira (1999): "Consumption and Portfolio Decisions When Expected Returns Are Time Varying," *Quarterly Journal of Economics*, 114, 433–496.

——— (2001a): *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*. Oxford Univeristy Press, London, UK.

——— (2001b): "Who Should Buy Long-Term Bonds?," *American Economic Review*, 91, 99–127.

Campbell, J. Y., and T. Vuolteenaho (2004): "Bad Beta, Good Beta," *American Economic Review*, 94, 1249–1275.

Campbell, J. Y., and M. Yogo (2006): "Efficient Tests of Stock Return Predictability," *Journal of Financial Economics*, 81, 27–60.

Carhart, M. M. (1997): "On the Persistence of Mutual Fund Performance," *Journal of Finance*, 52(1), 57–82.

Carpenter, J. (2000): "Does Option Compensation Increase Managerial Risk Appetite?," *Journal of Finance*, 55, 2311–2332.

Carroll, C. (2000): *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*. in Joel Slemrod, (ed), Oxford University Press, Harvard University Press: Cambridge, MA.

Carroll, C. D. (2006): "The Method of Endogenous Gridpoints for Solving Dynamic Stochastic Optimization Problems," *Economics Letters*, Forthcoming.

Chan, L. K., H.-L. Chen, and J. Lakonishok (2002): "On Mutual Fund Investment Styles," *Review of Financial Studies*, 15, 1407–1437.

Chan, L. K., S. G. Dimmock, and J. Lakonishok (2006): "Benchmarking Money Manager Performance: Issues and Evidence," Working Paper NBER No. 12461.

Chan, Y. L., and L. M. Viceira (2000): "Asset Allocation with Endogenous Labor Income: The Case of Incomplete Markets," Working Paper, Hong Kong University of Science and Technology and Harvard Business School.

Chang, W.-Y., Y.-N. Hsieh, and C.-C. Lai (2000): "Social Status, Inflation, and Endogenous Growth in a Cash-in-Advance Economy," *European Journal of Political Economy*, 16, 535–545.

Chapman, D. A., and Z. Xu (2007): "Career Concerns and the Active Fund Managers Problem," Working Paper, Boston College.

Charupat, N., and M. A. Milevsky (2002): "Optimal Asset Allocation in Life Annuities: A Note," *Insurance: Mathematics and Economics*, 30, 199–209.

Chen, H.-l., and G. G. Pennacchi (2007): "Does Prior Performance Affect a Mutual Fund's Choice of Risk? Theory and Further Empirical Evidence," Working Paper University of Illinois.

Chen, J., H. Hong, M. Huang, and J. D. Kubik (2004): "Does Fund Size Erode Mutual Fund Performance? The Role of Liquidity and Organization," *American Economic Review*, 94, 1276–1302.

Chen, L., and S. Zhao (2006): "Return Decomposition," Working Paper, Michigan State University and Kent State University.

Chetty, R. (2006): "A New Method of Estimating Risk Aversion," *American Economic Review*, 96, 1821–1834.

Chevalier, J., and G. Ellison (1997): "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 105, 1167–1200.

——— (1999a): "Are Some Mutual Fund Managers Better Than Others? Cross-sectional Patterns in Behavior and Performance," *Journal of Finance*, 54, 875–899.

Chevalier, J. A., and G. Ellison (1999b): "Career Concerns of Mutual Fund Managers," *Quarterly Journal of Economics*, 114, 389–432.

Cocco, J. (2005): "Portfolio Choice in the Presence Housing," *The Review of Financial Studies*, 18, 535–567.

Cocco, J. F., F. J. Gomes, and P. J. Maenhout (2005): "Consumption and Portfolio Choice over the Life-Cycle," *The Review of Financial Studies*, 18, 491–533.

Cochrane, J. H. (2006): "The Dog That Did Not Bark: A Defense of Return Predictability," Working Paper, University of Chicago.

Cochrane, J. H., and M. Piazzesi (2005): "Bond Risk Premia," *American Economic Review*, 95, 138–160.

——— (2006): "Decomposing the Yield Curve," Working Paper, University of Chicago.

Cohen, A., and L. Einav (2007): "Estimating Risk Preferences from Deductible Choice," *American Economic Review*, 97, 745–788.

Cohen, R., J. Coval, and L. Pástor (2005): "Judging Fund Managers by the Company They Keep," *Journal of Finance*, 60, 1057–1096.

Cole, H. L., G. J. Mailath, and A. Postlewaite (2001): "Investment and Concern for Relative Position," *Review of Economic Design*, 6, 241–261.

Cornell, B., and R. Roll (2005): "A Delegated-Agent Asset-Pricing Model," *Financial Analysts Journal*, 61, 57–69.

Cox, J., and C.-F. Huang (1989): "Optimal Consumption and Portfolio Policies when Asset Prices Follow a Diffusion Process," *Journal of Economic Theory*, 49, 33–83.

Cremers, M. (2002): "Stock Return Predictability: A Bayesian Model Selection Perspective," *Review of Financial Studies*, 15, 1223–1249.

Cremers, M., and A. Petajisto (2007): "How Active is Your Fund Manager? A New Measure That Predicts Performance," Working Paper Yale School of Management.

Cuoco, D., and R. Kaniel (2006): "Equilibrium Prices in the Presence of Delegated Portfolio Management," Working Paper Duke University and the University of Pennsylvania.

Cvitanic, J., and I. Karatzas (1992): "Convex Duality in Constrained Portfolio Optimization," *The Annals of Applied Probability*, 2, 767–818.

Cvitanic, J., A. Lazrak, L. Martellini, and F. Zapatero (2006): "Dynamic Portfolio Choice with Parameter Uncertainty and the Economic Value of Analysts' Recommendations," *Review of Financial Studies*, 19, 1113–1156.

Dai, Q., and K. J. Singleton (2000): "Specification Analysis of Affine Term Structure Models," *The Journal of Finance*, 50, 1943–1978.

——— (2002): "Expectation Puzzles, Time-varying Risk Premia, and Affine Models of the Term Structure," *Journal of Financial Economics*, 63, 415–441.

Dangl, T., Y. Wu, and J. Zechner (2007): "Market Discipline and Internal Governance in the Mutual Fund Industry," *Review of Financial Studies*, Forthcoming.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers (1997): "Measuring Mutual Fund Performance with Characteristic Based Benchmarks," *Journal of Finance*, 52, 1035–1058.

Davidoff, T., J. R. Brown, and P. A. Diamond (2005): "Annuities and Individual Welfare," *The American Economic Review*, 95, 1573–1590.

Davis, S. J., F. Kubler, and P. Willen (2003): "Borrowing Costs and the Demand for Equity over the Life Cycle," Working Paper.

Davis, S. J., and P. Willen (2000): "Using Financial Assets to Hedge Labor Income Risk: Estimating the Benefits," Working Paper.

de Jong, F. (2000): "Time-Series and Cross-Section Information in Affine Term Structure Models," *Journal of Business & Economic Statistics*, 18, 300–314.

De Marzo, P., R. Kaniel, and I. Kremer (2004): "Diversification as a Public Good: Community Effects in Portfolio Choice," *Journal of Finance*, 59, 1677–1715.

——— (2007): "Relative Wealth Concerns and Financial Bubbles," *Review of Financial Studies*, Forthcoming.

Deelstra, G., M. Grasselli, and P.-F. Koehl (2003): "Optimal Investment Strategies in the Presence of Labor of a Minimum Guarantee," *Insurance: Mathematics and Economics*, 33, 189–207.

Deli, D. (2002): "Mutual Fund Advisory Contracts: An Empirical Investigation," *Journal of Finance*, 57, 109–133.

Dessein, W., L. Garicano, and R. Gertner (2005): "Organizing for Synergies," Working paper, University of Chicago.

DeTemple, J. B., R. Garcia, and M. Rindisbacher (2003): "A Monte Carlo Method for Optimal Portfolios," *Journal of Finance*, 58, 401–446.

Diamond, P. A. (1997): "Macroeconomic Aspects of Social Security Reform," *Brookings papers on economic activity*, 2, 1–87.

Doucet, A., N. de Freitas, and N. Gordon (2001): *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.

Duffee, G. R. (2002): "Term Premia and Interest Rate Forecasts in Affine Models," *The Journal of Finance*, 57, 405–443.

Duffie, D., and R. Kan (1996): "A Yield Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Dunn, K. B., and J. J. McConnell (1981): "Valuation of Mortgage-Backed Securities," *The Journal of Finance*, 36, 599–617.

Durham, G. W., and A. R. Gallant (2002): "Numerical Techniques for Maximum Likelihood Estimation of Continuous-Time Diffusion Processes," *The Journal of Business and Economic Statistics*, 20, 297–316.

Dybvig, P. H., and L. Rogers (1997): "Recovery of Preferences from Observed Wealth in a Single Realization," *Review of Financial Studies*, 10, 151–174.

Edelen, R. M., R. B. Evans, and G. B. Kadlec (2007): "Scale Effects in Mutual Fund Performance: The Role of Trading Costs," Working Paper.

Eeckhout, J. (2004): "Gibrat's law for (All) Cities," *The American Economic Review*, 94, 1429–1451.

Eliasz, P. (2005): "Optimal Median Unbiased Estimation of Coefficients on Highly Persistent Regressors," Working Paper Princeton University.

Elton, E. J., and M. J. Gruber (2004): "Optimum Centralized Portfolio Construction with Decentralized Portfolio Management," *Journal of Financial and Quantitative Analysis*, 39, 481–494.

Elton, E. J., M. J. Gruber, and C. R. Blake (1996): "The Persistence of Risk-Adjusted Mutual Fund Performance," *Journal of Business*, 69, 133–157.

——— (2003): "Incentive Fees and Mutual Funds," *Journal of Finance*, 58, 779–804.

——— (2007): "Monthly Holdings Data and the Selection of Superior Mutual Funds," Working paper NYU Stern School of Business.

Epstein, L. G., and S. Zin (1989): "Substitution, Risk Aversion and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework," *Econometrica*, 57, 937–969.

Evans, R. (2007): "Does Alpha Really Matter? Evidence from Mutual Fund Incubation, Termination and Manager Change," Working Paper University of Virginia.

Fama, E. F. (2006): "The Behavior of Interest Rates," *Review of Financial Studies*, 19, 359–379.

Fama, E. F., and K. R. French (1988): "Dividend Yields and Expected Stock Returns," *Journal of Financial Economics*, 22, 3–27.

——— (1989): "Business Conditions and Expected Returns on Stocks and Bonds," *Journal of Financial Economics*, 25, 23–49.

Farhi, E., and S. Panageas (2005): "Saving and Investing for Early Retirement: A Theoretical Analysis," *Journal of Financial Economics, Forthcoming*.

Fernández-Villaverde, J., and J. Rubio-Ramírez (2004): "Estimating Dynamic Equilibrium Economies: Linear versus Nonlinear Likelihood," *Journal of Applied Econometrics*, 20, 891–910.

——— (2006): "Estimating Macroeconomic Models: A Likelihood Approach," *Review of Economic Studies,* forthcoming.

Fernández-Villaverde, J., J. Rubio-Ramírez, and M. Santos (2006): "Convergence Properties of the Likelihood of Computed Dynamic Models," *Econometrica*, 74, 93–119.

Ferson, W. E., S. Sarkissian, and T. T. Simin (2003): "Spurious regressions in financial economics?," *Journal of Finance*, 58(4), 1393–1413.

Finkelstein, A., and J. M. Poterba (2002): "Selection Effects in the United Kingdom Individual Annuities Market," *Economic Journal*, 112, 28–50.

Fontaine, J.-S., and R. Garcia (2007): "Bond Liquidity Premia," Unpublished paper, Universite de Montreal.

Foster, F. D., and M. Stutzer (2003): "Performance and Risk Aversion of Funds with Benchmarks: A Large Deviations Approach," Working paper, Australian Graduate School of Management.

Friedman, B. M., and M. J. Warshawsky (1990): "The Cost of Annuities: Implications for Saving Behavior and Bequests," *Quarterly Journal of Economics*, 105, 135–154.

Gabaix, X. (1999): "Zipf's Law for Cities: An Explanation," *Quarterly Journal of Economics*, 114, 739–767.

——— (2007): "Linearity-Generating Processes: A Modelling Tool Yielding Closed Forms for Asset Prices," Working Paper MIT.

Gabaix, X., and Y. Ioannides (2004): *The Evolution of City Size Distributions*. in V. Henderson and J-F. Thisse, (ed), Handbook of Regional and Urban Economics, North-Holland.

Gabaix, X., A. Krishnamurthy, and O. Vigneron (2006): "Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market," *The Journal of Finance*, Forthcoming.

Gertner, R. (1993): "Game Shows and Economic Behavior: Risk-Taking on 'Card Sharks'," *Quarterly Journal of Economics*, 108, 507–521.

Goel, A. M., and A. V. Thakor (2005): "Green with Envy: Implications for Corporate Investment Distortions," *Journal of Business*, 78, 2255–2287.

Goetzman, W. N., and P. Jorion (1995): "A Longer Look at Dividend Yields," *Journal of Business*, 68, 483–508.

Goffe, W. L., G. D. Ferrier, and J. Rogers (1994): "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, 60, 65–99.

Gollier, C., and J. W. Pratt (1996): "Risk Vulnerability and the Tempering Effect of Background Risk," *Econometrica*, 64, 1109–1124.

Gomes, F., and A. Michaelides (2005): "Optimal Life-Cycle Asset Allocation: Understanding the Empirical Evidence," *The Journal of Finance*, 60, 869–904.

Gomez, J.-P., and F. Zapatero (2003): "Asset Pricing Implications of Benchmarking: A Two-factor CAPM," *European Journal of Finance*, 9, 343–357.

Goriaev, A., T. E. Nijman, and B. J. Werker (2005): "Yet Another Look at Mutual Fund Tournaments," *Journal of Empirical Finance*, 12, 127–137.

Gourinchas, P.-O., and J. A. Parker (2002): "Consumption over the Life-Cycle," *Econometrica*, 70, 47–89.

Goyal, A., and I. Welch (2003): "Predicting the Equity Premium with Dividend Ratios," *Management Science*, 49(5), 639–654.

——— (2006): "A Comprehensive Look at The Empirical Performance of Equity Premium Prediction," *The Review of Financial Studies*, Forthcoming.

Greenwood, R., and D. Vayanos (2007): "Bons Supply and Excess Bond Returns," Working Paper, London School of Economics.

Gruber, M. J. (1996): "Another Puzzle: The Growth in Actively Managed Mutual Funds," *Journal of Finance*, 51, 783–810.

Haldane, J. (1942): "Moments of the Distributions of Powers and Products of Normal Variates," *Biometrika*, 32, 226–242.

Harvey, A. C. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

He, H., and N. Pearson (1991): "Consumption and Portfolio Policies with Incomplete Markets and Short-sale Constraints: The Infinite-dimensional Case," *Journal of Economic Theory*, 54, 259–304.

He, Z., and A. Krishnamurthy (2006): "Intermediation, Capital Immobility, and Asset Prices," Working paper Kellogg School of Management.

Heaton, J., and D. Lucas (1997): "Market Frictions, Savings Bahavior, and Portfolio Choice," *Macroeconomic Dynamics*, 1, 76–101.

——— (2000): "Portfolio Choice in the Presence of Background Risk," *The Economic Journal*, 110, 1–26.

Horneff, W. J., R. Maurer, O. S. Mitchell, and I. Dus (2006): "Optimizing the Retirement Portfolio: Asset Allocation, Annuitization, and Risk Aversion," Working paper, Pension Research Council, The Wharton School, University of Pennsylvania.

Hu, F., A. R. Hall, and C. Harvey (2000): "Promotion or Demotion? An Empirical Investigation of the Determinants of Top Mutual Fund Manager Change," Working Paper Duke University.

Hu, P., J. Kale, M. Pagani, and A. Subramaniam (2007): "Fund Flows, Performance, Managerial Career Concerns, and Risk-Taking: Theory and Evidence," Working paper Georgia State University.

Huberman, G. (2007): "Is the Price of Money Managers too Low?," Working paper Columbia University.

Hugonnier, J., and R. Kaniel (2007): "Mutual Fund Portfolio Choice in the Presence of Dynamic Flows," Working paper Fuqua School of Business.

Jamshidian, F. (1989): "An Exact Bond Option Formula," *Journal of Finance*, 44, 205–209.

Jarrow, R., and Y. Yildirim (2003): "Pricing Treasury Inflation Protected Securities and Related Derivatives Using an HJM Model," *Journal of Financial and Quantitative Analysis*, 38, 337–358.

Jazwinski, A. (1973): *Stochastic Processes and Filtering Theory*. Academic Press.

Jensen, M. (1968): "The Performance of Mutual Funds in the Period 1945-1964," *Journal of Finance*, 23, 389–416.

Jorion, P. (1996): "Dynamic Nonmyopic Portfolio Behavior," *Review of Financial Studies*, 9, 141–162.

——— (2003): "Portfolio Optimization with Tracking-error Constraints," *Financial Analysts Journal*, 59, 70–82.

Julier, S., and J. Uhlmann (1997): "A New Extension of the Kalman Filter to Nonlinear Systems," *Proceedings SPIE Signal Processing, Sensor Fusion, and Target Recognition VI*, 3068, 182–193.

Jullien, B., and B. Salanié (2000): "Estimating Preferences under Risk: The Case of Racetrack Bettors," *Journal of Political Economy*, 108, 503–530.

Jurek, J., and L. M. Viceira (2007): "Optimal Value and Growth Tilts in Long-Horizon Portfolios," Working Paper, Harvard University.

Kacperczyk, M., and A. Seru (2007): "Fund Manager Use of Public Information," *Journal of Finance*, Forthcoming.

Kacperczyk, M., C. Sialm, and L. Zheng (2005): "On the Industry Concentration of Actively Managed Equity Mutual Funds," *Journal of Finance*, 60, 1983–2011.

Kimball, M. S., C. R. Sahm, and M. D. Shapiro (2007): "Imputing Risk Tolerance From Survey Responses," Working paper NBER No. 13337.

Koijen, R. S. (2007): "Likelihood-based Estimation of Dynamic Models of Delegated Portfolio Management," Working Paper NYU Stern School of Business.

Koijen, R. S., and S. V. Nieuwerburgh (2007): "Financial Economics, Market Efficiency and Return Predictability," Unpublished paper, NYU Stern School of Business.

Koijen, R. S., T. E. Nijman, and B. J. Werker (2007a): "When Can Life-cycle Investors Benefit From Time-varying Bond Risk Premia?," Working Paper NYU Stern School of Business and Tilburg University.

Koijen, R. S. J., T. E. Nijman, and B. J. M. Werker (2007b): "Appendix Describing the Numerical Methods Used in 'When Can Life-cycle Investors Benefit from Time-varying Bond Risk Premia?'," Working Paper, Tilburg University.

——— (2007c): "Optimal Annuity Risk Management," Working Paper, Tilburg University.

Koijen, R. S. J., O. van Hemert, and S. van Nieuwerburgh (2007): "Mortgage Timing," Working Paper, Tilburg University and NYU Stern School of Business.

Kosowski, R., A. Timmermann, R. Wermers, and H. White (2006): "Can Mutual Fund "Stars" Really Pick Stocks? New Evidence from a Bootstrap Analysis," *Journal of Finance*, 56, 2551–2595.

Kothari, S., and J. Shanken (1992): "Stock Return Variation and Expected Dividends: A Time-series and Cross-sectional Analysis," *Journal of Financial Economics*, 31, 177–210.

Lamont, O. (1998): "Earnings and Expected Returns," *Journal of Finance*, 53, 1563–87.

Larrain, B., and M. Yogo (2007): "Does Firm Value Move Too Much to be Justified by Subsequent Changes in Cash Flow?," *Journal of Financial Economics,* forthcoming.

Lettau, M., and S. C. Ludvigson (2001): "Consumption, Aggregate Wealth and Expected Stock Returns," *Journal of Finance*, 56(3), 815–849.

——— (2005): "Expected Returns and Expected Dividend Growth," *Journal of Financial Economics*, 76, 583–626.

Lettau, M., and S. van Nieuwerburgh (2006): "Reconciling the Return Predictability Evidence," *The Review of Financial Studies*, Forthcoming.

Lewellen, J. W. (2004): "Predicting Returns With Financial Ratios," *Journal of Financial Economics*, 74, 209–235.

Liu, J. (2007): "Portfolio Selection in Stochastic Environments," *Review of Financial Studies*, 20, 1–39.

Liu, J., E. Peleg, and A. Subrahmanyam (2007): "Information, Expected Utility, and Portfolio Choice," Working paper UCLA.

Longstaff, F. A. (2005): "Borrower Credit and the Valuation of Mortgage-Backed Securities," *Real Estate Economics*, 33, 619661.

Lopes, P. (2005): "The Effects of Load Factors and Minimum Size Restrictions on Annuity Market Participation," Working paper, London School of Economics.

Lustig, H., and S. V. Nieuwerburgh (2005): "Housing Collateral, Consumption Insurance and Risk Premia: an Empirical Perspective," *Journal of Finance*, 60(3), 1167–1219.

Lustig, H., C. Syverson, and S. Van Nieuwerburgh (2007): "IT, Corporate Payouts, and Growing Inequality in Managerial Compensation," Working paper NYU Stern School of Business.

Lustig, H., and S. Van Nieuwerburgh (2006): "Can Housing Collateral Explain Long-Run Swings in Asset Returns?," Working Paper NYU Stern and UCLA.

Lustig, H., A. Verdelhan, and S. V. Nieuwerburgh (2007): "The Consumption-Wealth Ratio: A Litmus Test for Consumption-based Asset Pricing Models," Unpublished paper.

Luttmer, E. G. (2007): "Selection, Growth, and the Size Distribution of Firms," *Quarterly Journal of Economics*, Forthcoming.

Lynch, A. (2001): "Portfolio Choice and Equity Characteristics: Characterizing the Hedging Demands Induced by Return Predictability," *Journal of Financial Economics*, 62, 67–130.

Lynch, A., and P. Balduzzi (2000): "Predictability and Transaction Costs: The Impact on Rebalancing Rules and Behavior," *Journal of Finance*, 55, 2285–2310.

Lynch, A., and S. Tan (2006): "Labor Income Dynamics at Business-cycle Frequencies: Implications for Portfolio Choice," Working Paper, NYU Stern School of Business and Fordham University.

Lynch, A. W., and D. K. Musto (2003): "How Investors Interpret Past Fund Returns," *Journal of Finance*, 58, 2033–2058.

Lynch, A. W., and J. A. Wachter (2007a): "Does Mutual Fund Performance Vary over the Business Cycle?," Working Paper NYU Stern and The Wharton School.

——— (2007b): "Using Samples of Unequal Length in Generalized Method of Moments Estimation," Working Paper NYU Stern and The Wharton School.

Mamaysky, H., and M. Spiegel (2002): "A Theory of Mutual Funds: Optimal Fund Objectives and Industry Organization," Working paper Yale School of Management.

Massa, M., J. Reuter, and E. Zitzewitz (2007): "The Rise of Teams in Asset Management," Working Paper INSEAD.

Menzly, L., T. Santos, and P. Veronesi (2004): "Understanding Predictability," *Journal of Political Economy*, 112(1), 1–47.

Merton, R. C. (1973): "Theory of Rational Option Pricing," *Bell Journal of Economics and Management Science*, 4, 141–183.

——— (1980): "On Estimating the Expected Return on the Market," *Journal of Financial Economics*, 8, 323–361.

Metrick, A. (1995): "A Natural Experiment in 'Jeopardy!'," *American Economic Review*, 85, 240–253.

Milevsky, M. A., and V. R. Young (2003): "Annuitization and Asset Allocation," Working paper.

Mitchell, O. S., J. M. Poterba, M. J. Warshawsky, and J. R. Brown (1999): "New Evidence on the Money's Worth of Individual Annuities," *The American Economic Review*, 89, 1299–1318.

Munk, C., and C. Sørensen (2004): "Optimal Consumption and Investment Strategies with Stochastic Interest Rates," *Journal of Banking and Finance*, 28, 1987–2013.

Munk, C., and C. Sørensen (2005): "Dynamic Asset Allocation with Stochastic Income and Interest Rates," Working Paper.

Nanda, V., J. Wang, and L. Zheng (2007): "The ABCs of Mutual Funds: On the Introduction of Multiple Share Classes," Working Paper.

Nelson, C. C., and M. J. Kim (1993): "Predictable Stock Returns: The Role of Small Sample Bias," *Journal of Finance*, 43, 641–661.

Neuberger, A. (2003): "Annuities and the Optimal Investment Decision," Working paper, London Business School.

Nielsen, l. T., and M. Vassalou (2004): "Sharpe Ratios and Alphas in Continuous Time," *Journal of Financial and Quantitative Analysis*, Forthcoming.

Ou-Yang, H. (2003): "Optimal Contracts in a Continuous-time Delegated Portfolio Management Problem," *Review of Financial Studies*, 16, 173–208.

Panageas, S., and M. M. Westerfield (2007): "High Water Marks: High Risk Appetites? Convex Compensation, Long Horizons, and Portfolio Choice," *Journal of Finance*, Forthcoming.

Pástor, L. (2000): "Portfolio Selection and Asset Pricing Models," *Journal of Finance*, 55, 179–223.

Pástor, L., M. Sinha, and B. Swaminathan (2007): "Estimating the Intertemporal Risk-Return Tradeoff Using the Implied Cost of Capital," *Journal of Finance*, forthcoming.

Pástor, L., and R. F. Stambaugh (2002a): "Investing in Equity Mutual Funds," *Journal of Financial Economics*, 63, 351–380.

Pástor, L., and R. F. Stambaugh (2002b): "Mutual Fund Performance and Seemingly Unrelated Assets," *Journal of Financial Economics*, 63, 315–349.

——— (2006): "Predictive Systems: Living with Imperfect Predictors," Working Paper University of Chicago and CRSP.

Pástor, L., and P. Veronesi (2003): "Stock Valuation and Learning About Profitability," *Journal of Finance*, 58, 1749–1789.

——— (2006): "Was There a Nasdaq Bubble in the Late 1990s?," *Journal of Financial Economics*, 81, 61–100.

Piazzesi, M., M. Schneider, and S. Tuzel (2004): "Housing, Consumption, and Asset Pricing," Unpublished paper, University of Chicago.

Pliska, S. R. (2006): "Optimal Mortgage Refinancing with Endogenous Mortgage Rates: an Intensity Based Equilibrium Approach," Working Paper University of Illinois at Chicago.

Polk, C., S. Thompson, and T. Vuolteenaho (2006): "Cross-sectional Forecasts of the Equity Risk Premium," *Journal of Financial Economics*, 81, 101–141.

Pollet, J. M., and M. Wilson (2007): "How Does Size Affect Mutual Fund Behavior?," *Journal of Finance*, Forthcoming.

Poterba, J. M. (1997): "The History of Annuities in the United States," Working paper, NBER No. 6001.

Robson, A. J. (1992): "Status, the Distribution of Wealth, Private and Social Attitudes to Risk," *Econometrica*, 60, 837–857.

——— (2001): "The Biological Basis of Economic Bahavior," *Journal of Economic Literature*, 39, 11–33.

Roll, R. (1992): "A Mean-Variance Analysis of Tracking Error," *Quarterly Journal of Economics*, 18, 13–22.

Roussanov, N. (2007): "Diversification and its Discontents: Idiosyncratic and Entrepreneurial Risk in the Quest for Social Status," Working paper, The Wharton School.

Rytchkov, O. (2007): "Filtering Out Expected Dividends and Expected Returns," Unpublished paper, Texas A&M.

Sangvinatsos, A., and J. A. Wachter (2005): "Does the Failure of the Expectations Hypothesis Matter for Long-Term Investors?," *The Journal of Finance*, 60, 179–230.

Schwartz, A. (2007): "Household Refinancing Behavior in Fixed Rate Mortgages," Working Paper, Harvard University.

Schwartz, E. S., and W. N. Torous (1989): "Prepayment and the Valuation of Mortgage-Backed Securities," *The Journal of Finance*, 44, 375–392.

Sensoy, B. A. (2007): "Performance Evaluation and Self-Designated Benchmark Indices in the Mutual Fund Industry," Working Paper University of Southern California.

Sharpe, W. F. (1981): "Decentralized Investment Management," *Journal of Finance*, 36, 217–234.

——— (2002): "Budgeting and Monitoring Pension Fund Risk," *Financial Analysts Journal*, 58, 74–86.

Shen, P., and J. Corning (2001): "Can TIPS Help Identify Long-Term Inflation Expectations," *Economic Review, Federal Reserve Bank of Kansas City*, Fourth Quarter, 61–87.

Shore, S. H., and J. S. White (2006): "External Habit Formation and the Home Bias Puzzle," Working paper Wharton School of Business.

Sirri, E. E., and P. Tufano (1998): "Costly Search and Mutual Fund Flows," *Journal of Finance*, 53, 1589–1622.

Skinner, D. J. (2006): "The Evolving Relation between Earnings, Dividends, and Stock Repurchases," Working Paper University of Chicago.

Soares, C., and M. J. Warshawsky (2004): "Annuity Risk: Volatility and Inflation Exposure in Payments from Immediate Life Annuities," in *Developing an Annuity Market in Europe*, ed. by E. Fornero, and E. Lucian. Cheltenham: Edward Elgar.

St-Amour, P. (2006): "Benchmarks in Aggregate Household Portfolios," Working paper HEC University of Lausanne.

Stambaugh, R. F. (1999): "Predictive Regressions," *Journal of Financial Economics*, 54, 375–421.

Stanton, R. (1995): "Rational Prepayment and the Valuation of Mortgage-Backed Securities," *The Review of Financial Studies*, 8, 677–708.

Stanton, R., and N. Wallace (1998): "Mortgage Choice: What is the Point?," *Real Estate Economics*, 26(2), 173–205.

Storesletten, K., C. Telmer, and A. Yaron (2004): "Cyclical Dynamics in Idiosyncratic Labor-market Risk," *Journal of Political Economy*, 112, 695–717.

Stracca, L. (2006): "Delegated Portfolio Management: A Survey of the Theoretical Literature," *Journal of Economic Surveys*, 20, 823–848.

Stutzer, M. (2003a): "Fund Managers May Cause Their Benchmarks to Be Priced "Risk"," Working paper, University of Colorado.

——— (2003b): "Portfolio Choice with Endogenous Utility: A Large Deviations Approach," *Journal of Econometrics*, 116, 365–386.

Tauchen, G., and R. Hussey (1991): "Quadrature-based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models," *Econometrica*, 59, 371–396.

Torous, W., R. Valkanov, and S. Yan (2004): "On Predicting Returns with nearly Integrated Explanatory Variables," *Journal of Business*, 77, 937–966.

Treynor, J. L., and F. Black (1973): "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business*, 46, 66–86.

Valkanov, R. (2003): "Long-Horizon Regressions: Theoretical Results and Applications," *Journal of Financial Economics*, 68, 201–232.

Van Hemert, O. (2006): "Life-Cycle Housing and Portfolio Choice with Bond Markets," Working Paper, NYU Stern School of Business.

Van Nieuwerburgh, S., and L. Veldkamp (2006): "Information Acquisition and Portfolio Under-Diversification," Working Paper New York University.

——— (2007): "Information Immobility and the Home Bias Puzzle," Working paper NYU Stern School of Business.

Vayanos, D. (2003): "The Decentralization of Information Processing in the Presence of Interactions," *Review of Economic Studies*, 70, 667–695.

Viceira, L. (1996): "Testing For Structural Change in the Predictability of Asset Returns," Unpublished manuscript, Harvard University.

Viceira, L. M. (2001): "Optimal Portfolio Choice for Long-Horizon Investors with Non-Tradable Labor Income," *The Journal of Finance*, 56, 433–470.

Vickery, J. (2006): "Interest Rates and Consumer Choice in the Residential Mortgage Market," Working Paper, Federal Reserve Bank of New York.

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

Wachter, J. A. (2002): "Portfolio and Consumption Decisions Under Mean-Reverting Returns: An Exact Solution for Complete Markets," *Journal of Financial and Quantitative Analysis*, 37, 63–92.

——— (2003): "Risk Aversion and Allocation to Long-term Bonds," *Journal of Economic Theory*, 112, 325–333.

Wachter, J. A., and M. Warusawitharana (2007): "What is the Chance that the Equity Premium Varies over Time?," Working Paper, University of Pennsylvania.

Wan, E., and R. van der Merwe (2001): *The Unscented Kalman Filter*. Simon Haykin eds., Kalman Filtering and Neural Network, Wiley and Sons Publishing, New York.

Wermers, R., T. Yao, and J. Zhao (2007): "The Investment Value of Mutual Fund Portfolio Disclosure," Working paper Robert H. Smith School of Business.

Yaari, M. E. (1965): "Uncertain Lifetime, Life Insurance, and the Theory of the Consumer," *The Review of Economic Studies*, 32, 137–150.

Yuan, K. (2007): "Rank Fund Managers by the Accuracy of Their Beliefs," Working paper University of Michigan.

Zou, H. (1994): ""The Spirit of Capitalism" and Long-run Growth," *European Journal of Political Economy*, 10, 279–293.

——— (1995): "The Spirit of Capitalism and Savings Behavior," *Journal of Economic Behavior and Organization*, 28, 131–143.